Mesurer et comprendre le biais d'usage des codons : du recueil des applications à l'évolution des paralogues et des polyomavirus

Jérôme Bourret

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biologie de l'Évolution

Écoles doctorales CBS2 / GAIA

Unité de recherche : Maladies Infectieuses et Vecteurs : Écologie, Génétique, Évolution et Contrôle
(MIVEGEC ; UMR IRD 224-CNRS 5290-Université de Montpellier)

Mesurer et comprendre le biais d'usage des codons : du recueil des applications à l'évolution des paralogues et des polyomavirus.

Présentée par Jérôme BOURRET
Le 14 décembre 2020

Sous la direction de
Ignacio González BRAVO et Samuel ALIZON

Devant le jury composé de

Celine SCORNAVACCA, Directrice de Recherche, Université de Montpellier	Présidente
Gwenaël PIGANEAU, Directrice de Recherche, CNRS, Banyuls-sur-Mer	Rapporteure
Laurent DURET, Directeur de Recherche, CNRS, Villeurbanne	Rapporteur
Céline BRESSOLLETTE, Maîtresse de Conférences Praticienne Hospitalière, Université de Nantes	Examinatrice
Anna-Sophie FISTON-LAVIER, Maîtresse de Conférences, Université de Montpellier	Examinatrice
Ignacio González BRAVO, Directeur de Recherche, CNRS, Montpellier	Directeur de thèse
Samuel ALIZON, Directeur de Recherche, CNRS, Montpellier	Directeur de thèse

UNIVERSITÉ DE MONTPELLIER

TABLE DES MATIÈRES

Liste des Figures

Liste des Tables

Mesurer et comprendre le biais d'usage des codons : recueil des applications à l'évolution des paralogues et des polyomavirus.

CADRE DU DOCTORAT

1 Préface

Ce doctorat a été réalisé en trois ans dans le cadre d'un financement public partagé entre les écoles doctorales CBS2 et GAIA de l'Université de Montpellier. Il s'insère au sein des thématiques des Laboratoires MIVEGEC – Maladies Infectieuses et Vecteurs : Écologie, Génétique, Évolution et Contrôle – (UMR IRD 224-CNRS 5290-Université de Montpellier), qui se définissent par l'étude transdisciplinaire des mécanismes de maintenance, d'amplification et de transmission d'agents pathogènes, de leurs déterminants génétiques et non génétiques, et ce afin de pouvoir mieux appréhender l'évolution des systèmes infectieux et contribuer à en améliorer le contrôle. Précisément, ce projet a été majoritairement encadré par Ignacio G. Bravo, mais aussi par Samuel Alizon, au sein des équipes VIROSTYLE et ETE, qui se penchent sur l'évolution virale et sur la diversité des styles de vie des virus, avec comme terrain d'étude principal l'étude des papillomavirus et des maladies oncogéniques qu'ils peuvent provoquer.

L'une des thématiques de l'équipe VIROSTYLE est l'étude du biais d'usage des codons dans le cadre d'une relation hôte-pathogène entre un virus et son hôte. Cette thèse se focalise donc sur la nature du biais d'usage des codons, sa mesure, son analyse et ses particularités dans l'horizon des organismes pour, *in fine*, mieux comprendre le rôle du biais d'usage des codons dans la relation entre un hôte (*Homo sapiens*) et les virus qui l'infectent. Pour illustrer ce phénomène, VIROSTYLE se penche notamment sur l'histoire évolutive et sur les spécificités d'usage des codons des polyomavirus. L'un des projets annexes de cette étude, en partenariat avec l'ANR BK-NAB du CHU de Nantes, se focalise sur la dynamique virale des polyomavirus BK pendant l'évolution clinique d'une greffe de rein vers une *PolyomaVirus-Associated Nephropathy* (PVAN). L'objectif de cette étude, dont les données sont toujours en cours d'obtention, est d'analyser l'évolution des profils génétiques du polyomavirus BK au cours d'une infection chez un patient immunosupprimé, et ce pour mieux comprendre le fondement de la prolifération virale et des mécanismes d'échappement de l'immunité de l'hôte conduisant à une PVAN.

Au sein de cette thèse, notre travail de recherche s'est articulé autour de ces deux principales thématiques qui sont a priori distinctes mais qui trouveront un lien au cours des années suivantes. Dans un premier temps, nous explorerons les biais d'usage des codons sous des perspectives mathématiques, informatiques et analytiques. Par la suite, nous présenterons les travaux préliminaires que nous avons effectués sur l'analyse des polyomavirus BK.

2 Analyse du biais d'usage des codons

Depuis plusieurs décennies, le biais d'usage des codons (BUC ; CUB ou CUPrefs pour *Codons Usage Preferences* en anglais) est étudié sous de multiples aspects, et ce dans l'objectif de

délier ses possibles effets sur la traduction des biais mutationnels et de composition nucléotidique qui peuvent le façonner. Sans être contradictoires, les deux grandes hypothèses sélectionnistes et neutralistes pour expliquer l'origine du CUB sont plus ou moins acceptées selon l'espèce considérée. C'est notamment chez les Vertébrés que les forces à l'origine du CUB sont débattues : la plupart des auteurs s'accordent à souligner une prédominance des biais mutationnels, sans qu'aucune preuve formelle d'une sélection traductionnelle n'ait pu être mise en avant. Par ailleurs, la mesure du CUB manque d'outils mathématiques et informatiques. Celle-ci est principalement assurée par les indices CAI et ENC, non exempts de défauts : l'indice CAI est souvent considéré à tort comme un indice de mesure du CUB (alors qu'il mesure l'« optimalité » d'une séquence par rapport à un jeu de référence) et l'ENC a pour principal défaut d'être grandement dépendant de la taille et de la composition de la séquence requête.

Au sein de cette thèse, nous approfondissons l'analyse du CUB de ces organismes d'un point de vue mathématique, informatique et analytique. Nous développons ainsi COUSIN, un nouvel indice comparatif du biais d'usage des codons et dont les résultats sont normalisés autour d'une Hypothèse Nulle (H_0) décrivant un usage égal des codons synonymes. En parallèle du développement de cet indice, nous avons créé `COUSIN`, un programme éponyme simple d'utilisation et proposant une analyse poussée du biais d'usage des codons. Cette étude a fait l'objet d'une publication au sein du journal Genome and Biology Evolution (GBE, éditions Oxford) sous le PMID 31800035. Nous étrennons notre nouvel indice au sein d'une étude sur les gènes paralogues *PTBP*.

Les gènes paralogues *PTBP*, présents chez tous les Vertébrés, exercent la même fonction mais affichent une différence marquée au niveau de leur CUB et de leur expression au sein des différents tissus. Plus encore, un des paralogues voit son expression augmenter *in vitro* si l'on modifie son CUB. Dans l'objectif de déterminer un possible accord entre évolution, CUB et expression génique, nous analysons l'histoire évolutive des *PTBP* et la confrontons à leur biais d'usage des codons. Ces travaux ont été soumis à GBE et peuvent être consultés sur la plateforme BioRxiv suivant ce lien hypertexte : https://doi.org/10.1101/2020.08.30.274191.

3 Évolution des polyomavirus

Les polyomavirus, appartenant à la famille des *Polyomaviridae*, constituent un taxon infectant une large gamme d'hôtes allant des arthropodes aux mammifères. Il existe à ce jour quinze polyomavirus humains, non monophylétiques, divergeants et ayant différents tropismes cellulaires. Bien que ces virus provoquent la plupart du temps des infections asymptomatiques, certains d'entre eux peuvent être occasionnellement virulents à différents degrés chez une personne immunosupprimée. À titre d'exemple, le polyomavirus JC peut provoquer une leucoencéphalopathie multifocale progressive, le polyomavirus à cellules de Merkel un cancer de la peau et le polyomavirus BK, dans le cadre d'une greffe de reins, peut mener à une défaillance de l'organe par le développement d'une PVAN. Chez le polyomavirus BK, les mécanismes d'activation de la pathogénicité et de sa gravité sont encore méconnus. Il est néanmoins estimé que l'immunosuppression dans le cadre d'une greffe profiterait à une prolifération et à une évolution intra-patient menant à l'apparition de souches virulentes qui provoqueraient l'inflammation du greffon et sa

perte de fonction. Le seul traitement existant, à ce jour, est une modulation de l'immunosuppression pouvant elle-même mener à un rejet de la greffe par le système immunitaire. Dans l'espoir de mieux comprendre les mécanismes évolutifs intra-patient des polyomavirus BK, nous effectuons une analyse de leur histoire évolutive et de leur profil génétique au cours d'une infection chez un receveur de greffe de rein.

Une première partie de ce projet se focalise sur la création de nouvelles méthodes pour génotyper et analyser les génomes des polyomavirus humains, et ce dans une optique de faciliter les travaux cliniques sur ces virus. Nous avons développé pour cela deux outils informatiques :

— `ViroPhylo`, une plate-forme web permettant de génotyper facilement et avec robustesse les polyomavirus humains. Il s'agit d'un *pipeline* (*i.e.* une chaîne de traitement) se basant sur l'insertion de séquences requêtes au sein d'une phylogénie de référence.

— `GenoPolys`, un *pipeline* permettant d'analyser avec précision des génomes complets de polyomavirus BK.

Par ailleurs, nous analysons l'histoire évolutive des polyomavirus humains, et explorons la diversité en CUB qu'ils présentent, aussi bien au niveau des gènes précoces que tardifs. Le CUB des polyomavirus humains ne suit pas leur phylogénie, et pourrait être marqué par des signatures de sélection. Nous en profitons pour ouvrir de nouveau le débat sur le CUB des Vertébrés par le biais des virus qui les infectent.

Chapitre Un 1

INTRODUCTION

1.1 Transcription, traduction et codons synonymes

1.1.1 Rappels sur les processus de transcription et de traduction

Dans les domaines de la biologie cellulaire et moléculaire, la transcription d'un gène et la traduction de sa séquence codante (SC ; CDS en anglais) décrivent le passage de sa forme génomique vers sa forme transcrite puis protéique. Ces deux processus sont à l'origine de l'expression des gènes et sont au cœur de la machinerie cellulaire [2].

La transcription d'un gène est assurée par le complexe moléculaire de l'ARN polymérase [3]. Celle-ci débouche sur la formation d'un ARN messager (ou ARNm), une copie éphémère et monocaténaire du gène à partir du brin d'ADN dit transcrit (Figure 1.1). Pour initier ou réguler la transcription, un certain nombre de protéines, appelées facteurs de transcription, sont nécessaires [4]. À titre d'exemple, les protéines de liaison à l'ADN telles que celles de la famille STAT (pour *Signal Transducers and Activators of Transcription*) se fixent à la molécule d'ADN pour la rendre plus ou moins accessible à l'ARN polymérase [5]. Une fois la transcription amorcée par les facteurs associés, l'ARN polymérase se fixe sur la région promotrice du gène et synthétise l'ARNm jusqu'à la rencontre de facteurs de terminaison de la transcription. Ces facteurs sont, par exemple, les terminateurs Rho-indépendants et Rho-dépendants des procaryotes, ou encore les mécanismes conduisant à la polyadénylation des régions 3' des ARNm d'eucaryotes [6, 7]. Chez les procaryotes, cet ARNm est directement traduit en une protéine (et bien souvent en parallèle de la transcription du gène) mais chez les eucaryotes, l'ARN est dit pré-messager et doit subir plusieurs étapes de maturation ainsi qu'une sortie du noyau avant d'être traduit en une protéine [8, 9].

Pour plusieurs raisons liées à leur activité au sein de la cellule, les ARNm se dégradent naturellement à une vitesse bien plus élevée que l'ADN. Leur demi-vie peut varier de quelques minutes (comme l'ARNm de la protéine ¸fos) [10] à environ une journée (ARNm de la protéine de l'ovalbumine) [11], mais leur longévité moyenne est de quelques heures [12]. Même s'il existe des exceptions, on peut considérer que plus la durée de vie d'un ARNm est longue, plus le gène associé sera exprimé. La structure secondaire d'un ARNm est un élément primordial dans sa longévité : elle correspond à la conformation que va prendre l'ARNm par le biais des interactions

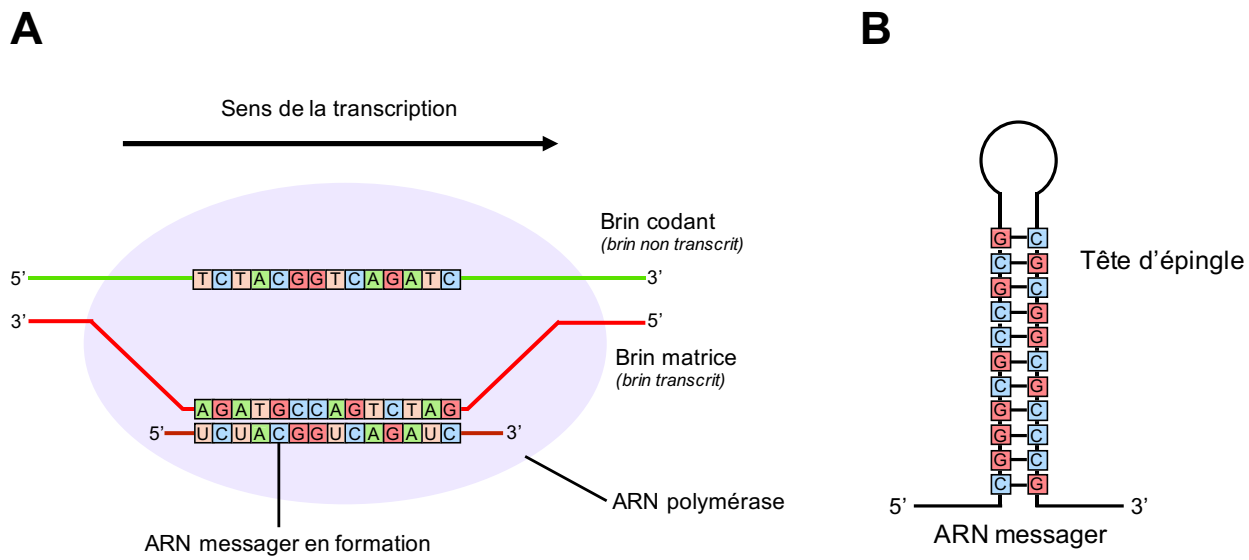

Figure 1.1 – **Mécanisme de transcription de l'ADN en un ARN messager (A) et structure secondaire en tête d'épingle (B).** A) La transcription est effectuée par l'ARN polymérase sur le brin transcrit de l'ADN (ou brin matrice) dans son sens 3'-5'. Chez les bactéries, il n'existe qu'une seule ARN polymérase, alors que plusieurs types d'ARN polymérases coexistent au sein des cellules eucaryotes. Chez ces dernières, seule l'ARN polymérase II est responsable de la transcription des ARNm. L'ARN polymérase synthétise la séquence en prenant comme motif le brin transcrit et L'ARNm qui en résulte est donc une copie monocaténaire du brin non-transcrit (ou brin codant). B) La structure secondaire des ARNm implique notamment la formation de séquences bicaténaires comme la tête d'épingle décrite au sein de cette figure. Dans ce cas, la tête d'épingle est formée par appariement de nucléotides G et C le long de la région bicaténaire.

entre les différents nucléotides qui le composent. Typiquement, des régions répétées inversées peuvent conduire à la formation de « têtes d'épingles », où l'ARNm devient localement bicaténaire. Une modification de la composition nucléotidique d'un ARNm peut entraîner un changement de sa structure secondaire et donc de sa stabilité [13, 14]. Plusieurs autres facteurs peuvent entraîner le raccourcissement ou le rallongement de la durée de vie d'un ARNm. Par exemple, les ARNm les plus éphémères possèdent généralement sur leur région 3' une courte séquence non codante riche en nucléotides AU (Adénine ou Uracile) favorable à l'action dégradante des exosomes [15]. Dans un autre cas, la présence de cytokine IL-4 au cours d'une inflammation a pour effet d'augmenter la longévité de l'ARNm de VCAM-1 pour augmenter son expression au sein des cellules endothéliales. La demi-vie d'un ARNm peut donc être sujette à une modulation selon son environnement [16]. La stabilité d'un ARNm est donc assurée par plusieurs facteurs aussi bien mécaniques que biochimiques, et on peut considérer qu'elle détermine grandement son expressivité au sein de la cellule ; la modulation de la demi-vie de ces séquences, et donc de leur temps d'accessibilité à la machinerie ribosomale, pourrait être un facteur important de leur traduction.

La traduction de l'ARNm en une protéine est assurée par la machinerie ribosomale, elle-même constituée de complexes ribonucléoprotéiques [17–19]. Les ribosomes sont composés de deux sous-unités définies par leurs coefficients de sédimentation (mesurés en *Svedberg*, de symbole S) : une petite (30S chez les procaryotes, 40S chez les eucaryotes) et une grande (50S chez les procaryotes, 60S chez les eucaryotes) [19, 20]. Lors de l'étape d'initiation de la traduction, les deux sous-unités du ribosome s'assemblent au niveau des premiers nucléotides de la région 5' de l'ARNm par le consort de différents facteurs d'initiation de la traduction [21]. Le ribosome débute alors sa traduction par la lecture successive de triplets de nucléotides appelés codons, et ce dans le sens 5'-3' de l'ARNm (Figure 1.2) [22, 23]. De par l'univers des nucléotides retrouvés au sein d'un ARNm (*i.e.*, les nucléotides « A » pour Adénine, « U » pour Uracile, « G » pour Guanine, « C » pour Cytosine), il existe 4^3 – soit 64 – combinaisons de codons. Chez la plupart des eucaryotes, dont le code génétique nucléaire est dit « standard », trois codons sont terminateurs de la traduction et sont communément appelés codons STOP. La lecture des autres codons conduit à l'appel par le ribosome de structures ribonucléiques nommées ARN de transfert (ou ARNt) [2] (Tableau 1.1). Ces ARNt contiennent notamment une courte séquence trinucléotidique complémentaire d'un codon (appelée l'anticodon) et sont chargés d'un acide aminé qui constitue l'unité de base d'une protéine. Le ribosome possède trois sites successifs d'accueil des ARNt : A (pour « Acide aminé », site d'entrée de l'ARNt chargé), P (pour « Peptide », site occupé par l'ARNt lié à la chaîne polypeptidique en cours de synthèse) et E (Pour « Exit », site de sortie de l'ARNt déchargé) [17]. Au cours de l'étape d'élongation de la traduction, les ARNt s'insèrent au sein du site A et sont confrontés aux codons lus par le ribosome. S'il y a correspondance entre le codon et l'anticodon, le ribosome se décale de manière à ce que le complexe codon-anticodon se retrouve sur le site P. L'acide aminé est alors polymérisé à la chaîne polypeptidique de la protéine en cours de formation. L'entrée d'un nouvel ARNt sur le site A entraîne le passage de l'ARNt déchargé du site P au site E. Cet ARNt déchargé est finalement évacué de la machinerie ribosomale lors de l'arrivée d'un troisième ARNt sur le site A (Figure 1.2). S'il n'y a pas de correspondance entre l'anticodon de l'ARNt et le codon lu, l'ARNt sort du site A et la machinerie ribosomale

en appelle un nouveau. La traduction s'arrête lorsque le ribosome rencontre un codon STOP qui engendrera son démantèlement par l'action de facteurs de terminaison de la traduction [17]. Par simple principe de continuité entre ADN, ARN et protéine, il est souvent considéré que le nombre de protéines au sein d'une cellule à un temps t est directement proportionnel au nombre d'ARNm présents. En réalité, la variation des quantités d'ARNm n'expliquerait qu'environ 40 % des variations des quantités de protéines, et ce pour diverses raisons dont certaines seront décrites au sein de cette thèse [24, 25].

Chez les procaryotes, la traduction est effectuée au sein du cytoplasme, alors que les protéines eucaryotes sont formées principalement au sein du réticulum endosplasmique granuleux [8, 9]. Une fois synthétisées, les protéines doivent subir une étape de maturation. Elle doivent se replier correctement et parfois être l'objet de modifications biochimiques pour acquérir leur fonction. Les protéines chaperonnes sont réputées pour aider au repliement correct des protéines [26].

Ce sont donc 61 codons qui sont théoriquement associés à un nombre égal d'anticodons qui leur sont complémentaires (Tableau 1.1). Mais en réalité, les associations codons-anticodons sont soumises au *wobble-effect* (ou *wobble pairing*, paire oscillante en français), permettant l'acceptation de plusieurs codons pour un même anticodon, la plupart du temps par un permissivité sur la troisième base [27, 28] (Figure 1.2). Ce type d'association entre codons et anticodons s'oppose aux associations canoniques dites de Watson-Crick, définies par les paires A:U et G:C [27]. Les ARNt à anticodon GNN (N représentant un des quatre nucléotide A, U, G ou C), normalement associés avec le codon NNC peuvent, par le biais de ce phénomène, s'associer par exemple avec les codons NNU. De par cette particularité, certains organismes comme *Escherichia coli* ou encore *H. sapiens* possèdent des codons pour lesquels il n'existe pas d'ARNt associés dans leur génome et donc d'anticodons canoniques complémentaires (Tableau 1.2). L'association se fait alors exclusivement avec un anticodon d'un ARNt proche que l'on appelle *near-cognate* (presque apparenté en français, par opposition à *cognate* ou apparenté en français) [29].

Outre la prise en compte du *wobble-effect* dans la traduction, plusieurs anticodons (et donc codons) sont associés à un même acide aminé. Ces codons homologues sont dits « synonymes » en référence à la dégénérescence du code génétique (Tableau 1.1) [23]. Au sein du code génétique standard, seules la méthionine et la tryptophane sont codées par un seul et unique codon, alors que tous les autres acides aminés sont codés par deux à six codons synonymes. Les ARNt susceptibles d'être chargés d'un même acide aminé sont quant à eux appelés iso-accepteurs [30].

1.1.2 Biais d'usage des codons

De par la dégénérescence du code génétique, on pourrait s'attendre à un usage équivalent des codons synonymes au sein des régions codantes d'un organisme. Mais, au sein d'un génome, d'une région génomique ou d'un gène, l'usage des codons synonymes de chaque acide aminé est systématiquement soumis à un déséquilibre qui, a priori, n'apporte aucun changement dans le phénotype des protéines obtenues [31]. Le Biais d'Usage des Codons (BUC ; en anglais CUB ou encore CUPrefs pour *Codon Usage Preferences*) se définit comme un usage différentiel des codons synonymes et ce à l'échelle d'un gène, d'une région génomique ou d'un génome entier [32, 33] (Figure 1.3 ; Tableau 1.2). À titre d'exemple, le CUB de la phénylalanine du gène

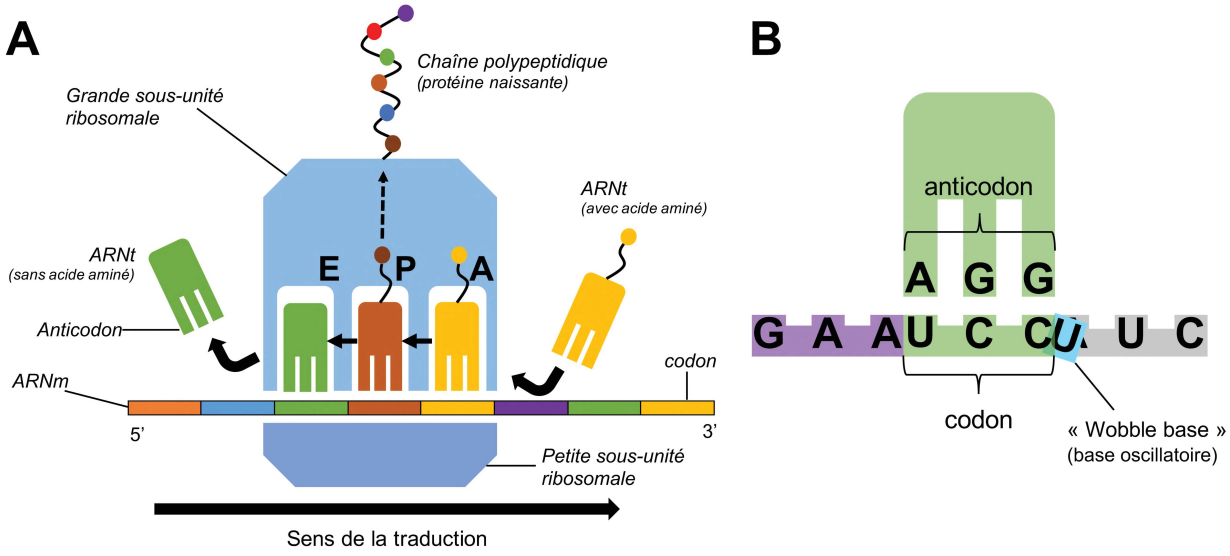

Figure 1.2 – **Mécanisme de traduction d'un ARN messager (A) et de *wobbling-effect* pouvant subvenir lors de la traduction (B).** A) L'initiation de la traduction se fait par l'assemblage du ribosome sur l'ARNm transcrit et mature. La phase d'élongation est représentée au sein de cette figure : le ribosome glisse le long de l'ARNm dans le sens 5'-3' de l'ARNm et appelle des ARNt au sein de son site A à chaque rencontre d'un nouveau codon. S'il y a complémentarité entre les deux séquences nucléotidiques de l'ARNm et de l'ARNt, le ribosome se décale d'un codon et l'ARNt sélectionné se place au sein du site P. L'acide aminé rejoint alors la chaîne polypeptidique en formation. Une fois cette étape effectuée, et l'insertion d'un nouvel ARNt sur le site A, l'ARNt déchargé est déplacé vers le site E. Il sera éjecté de la machinerie ribosomale lors d'un prochain décalage. B) Le wobbling-pairing, ou wobbling-effect, décrit un phénomène pouvant survenir lors de l'appariement entre le codon lu et l'anticodon de l'ARNt au sein du site A du ribosome. Dans cet exemple, l'anticodon AGG peut s'apparier avec les codons UCC et UCU. Cet ARNt, qui peut donc s'associer avec plusieurs codons, est qualifié *near-cognate*.

TABLE 1.1 – **Table inversée des codons retrouvés au sein du code génétique standard.** Les acides aminés sont organisés selon un ordre alphabétique et numérique suivant le nombre de codons.

Acide aminé	**Codons**
Met	AUG
Trp	UGG
Asn	AAU, AAC
Asp	GAU, GAC
Cys	UGU, UGC
Gln	CAA, CAG
Glu	GAA, GAG
His	CAU, CAC
Lys	AAA, AAG
Phe	UUU, UUC
Tyr	UAU, UAC
Ile	AUU, AUC, AUA
Ala	GCU, GCC, GCA, GCG
Gly	GGU, GGC, GGA, GGG
Pro	CCU, CCC, CCA, CCG
Thr	ACU, ACC, ACA, ACG
Val	GUU, GUC, GUA, GUG
Arg	CGU, CGC, CGA, CGG, AGA, AGG
Leu	UUA, UUG, CUU, CUC, CUA, CUG
Ser	UCU, UCC, UCA, UCG, AGU, AGC
$Terminateurs\ (codons\ STOP)$	UAG, UAA, UGA

TABLE 1.2 – **Tableau récapitulatif du CUB et du nombre de copies de gènes d'ARNt chez *E. coli*//*H.sapiens*.** Le premier trinucléotide indique le codon (dans le sens 5'-3'), et ses fréquences au sein des génomes de *E. coli* et *H. sapiens* sont données entre parenthèses. Le second trinucléotidique indique l'anticodon (dans le sens 5'-3'), et le nombre de copies d'ARNt retrouvées au sein des génomes *E. coli* et *H. sapiens* est indiqué entre parenthèses. Le dernier caractère de chaque cellule indique l'acide aminé associé au codon et à l'anticodon. Ces données sont issues de la base de données GtRNAdb [40]

UUU (0.57 / 0.47) | **AAA** (0 / 0) | F	**UCU** (0.15 / 0.19) | **AGA** (0 / 9) | S	**UAU** (0.57 / 0.45) | **AUA** (0 / 0) | Y	**UGU** (0.45 / 0.47) | **ACA** (0 / 0) | C
UUC (0.43 / 0.53) | **GAA** (2 / 10) | F	**UCC** (0.15 / 0.21) | **GGA** (2 / 0) | S	**UAC** (0.43 / 0.55) | **GUA** (3 / 13) | Y	**UGC** (0.55 / 0.53) | **CGA** (1 / 29) | C
UUA (0.13 / 0.08) | **UAA** (1 / 4) | L	**UCA** (0.12 / 0.15) | **UGA** (1 / 4) | S	**UAA** | **UUA** | STOP	**UGA** | **UCA** | STOP
UUG (0.13 / 0.13) | **CAA** (1 / 6) | L	**UCG** (0.15 / 0.06) | **CGA** (1 / 4) | S	**UAG** | **CUA** | STOP	**UGG** (1 / 1) | **CCA** (1 / 7) | W
CUU (0.10 / 0.13) | **AAG** (0 / 9) | L	**CCU** (0.16 / 0.29) | **AGG** (0 / 9) | P	**CAU** (0.57 / 0.43) | **AUG** (0 / 0) | H	**CGU** (0.39 / 0.08) | **ACG** (4 / 7) | R
CUC (0.10 / 0.19) | **GAG** (1 / 0) | L	**CCC** (0.12 / 0.32) | **GGG** (1 / 0) | P	**CAC** (0.43 / 0.57) | **GUG** (1 / 9) | H	**CGC** (0.40 / 0.18) | **GCG** (0 / 0) | R
CUA (0.04 / 0.07) | **UAG** (1 / 3) | L	**CCA** (0.19 / 0.28) | **UGG** (1 / 7) | P	**CAA** (0.35 / 0.27) | **UUG** (2 / 6) | Q	**CGA** (0.06 / 0.11) | **UCG** (0 / 6) | R
CUG (0.50 / 0.39) | **CAG** (4 / 9) | L	**CCG** (0.54 / 0.12) | **CGG** (1 / 4) | P	**CAG** (0.66 / 0.73) | **CUG** (2 / 13) | Q	**CGG** (0.09 / 0.20) | **CCG** (1 / 4) | R
AUU (0.51 / 0.37) | **AAU** (0 / 15) | I	**ACU** (0.17 / 0.25) | **AGU** (0 / 9) | T	**AAU** (0.45 / 0.48) | **AUU** (0 / 0) | N	**AGU** (0.15 / 0.15) | **ACU** (0 / 0) | S
AUC (0.42 / 0.46) | **GAU** (3 / 3) | I	**ACC** (0.44 / 0.34) | **GGU** (0 / 26) | T	**AAC** (0.55 / 0.51) | **GUU** (4 / 25) | N	**AGC** (0.28 / 0.24) | **GCU** (1 / 8) | S
AUA (0.07 / 0.18) | **UAU** (0 / 5) | I	**ACA** (0.13 / 0.29) | **UGU** (1 / 6) | T	**AAA** (0.77 / 0.44) | **UUU** (6 / 12) | K	**AGA** (0.03 / 0.22) | **UCU** (1 / 6) | R
AUG (1 / 1) | **CAU** (6 / 20) | M	**ACG** (0.27 / 0.11) | **CGU** (2 / 5) | T	**AAG** (0.23 / 0.56) | **CUU** (0 / 15) | K	**AGG** (0.02 / 0.21) | **CCU** (1 / 5) | R
GUU (0.26 / 0.19) | **AAC** (0 / 9) | V	**GCU** (0.16 / 0.26) | **AGC** (0 / 26) | A	**GAU** (0.63 / 0.47) | **AUC** (0 / 0) | D	**GGU** (0.34 / 0.13) | **ACC** (0 / 0) | G
GUC (0.21 / 0.24) | **GAC** (2 / 0) | V	**GCC** (0.27 / 0.40) | **GGC** (2 / 0) | A	**GAC** (0.38 / 0.53) | **GUC** (3 / 13) | D	**GGC** (0.41 / 0.34) | **GCC** (4 / 14) | G
GUA (0.15 / 0.12) | **UAC** (5 / 5) | V	**GCA** (0.21 / 0.230) | **UGC** (3 / 8) | A	**GAA** (0.69 / 0.43) | **UUC** (4 / 8) | E	**GGA** (0.10 / 0.25) | **UCC** (1 / 9) | G
GUG (0.37 / 0.46) | **CAC** (0 / 13) | V	**GCG** (0.36 / 0.111) | **CGC** (0 / 4) | A	**GAG** (0.31 / 0.57) | **CUC** (0 / 8) | E	**GGG** (0..15 / 0.25) | **CCC** (1 / 5) | G

Figure 1.3 – **Illustration du biais d'usage des codons d'une séquence.** Les codons synonymes AAU (rouge) et AAC (bleu) codent l'acide aminé asparagine. Au sein de cet exemple, le codon AAU est utilisé quatre fois sur cinq pour coder l'asparagine alors que le codon AAC est utilisé une fois sur cinq.

bactérien *dnaA* tend vers une surutilisation du codon UUC par rapport au codon UUU chez *Streptomyces coelicolor* (UUC:UUU avec un ratio de 0.94:0.06), alors que son CUB est plus équilibré chez *E. coli* (UUC:UUU avec un ratio de 0.57:0.43) (observations personnelles). Il est couramment admis que le CUB puise son origine dans deux propriétés non-exclusives. La première, appelée biais mutationnel, correspond aux processus mutationnels façonnant la composition nucléotidique d'une région d'intérêt vers un contenu riche en certains nucléotides, modifiant alors les fréquences des codons synonymes des gènes contenus dans ladite région [34–36]. La seconde, dénommée sélection traductionnelle, consiste en la sélection de codons synonymes permettant une meilleure expression du gène, et ce en considération des mécanismes relatifs à la transcription et à la traduction [37–39]. Ces deux origines du CUB seront débattues au sein de la prochaine section.

1.2 État des connaissances sur le biais d'usage des codons

Cela fait environ cinq décennies que le biais d'usage des codons est étudié sous de multiples aspects [41]. Peu de temps après la découverte du mécanisme de traduction, où l'on supposait un usage équivalent des codons synonymes, le séquençage toujours plus important de gènes viraux, bactériens et d'eucaryotes invita rapidement à une prise de conscience d'un déséquilibre quasi-constant des codons synonymes [32, 41, 42]. Plusieurs hypothèses neutralistes – où le CUB est influencé par la composition nucléotidique de la région génomique qui l'entoure – et sélectionnistes – où le CUB est déterminé comme jouant un rôle dans l'expression des gènes – ont vu le jour et continuent d'animer les passions [36, 37, 43, 44]. Au sein de ces discussions, l'un des dogmes les plus disputés est l'effet du CUB sur l'efficacité de la traduction et son influence sur le fonctionnement de la machinerie ribosomale [45, 46]. Au cours de cette section, nous tenterons d'englober les principales hypothèses associées au CUB au sein de différents organismes. Nous attirons l'attention du lecteur sur le terme, parfois controversé, d'« optimisation » des codons qui est couramment utilisé au sein des études sur le CUB. Cette terminologie sous-entend une adhérence aux hypothèses sélectionnistes (de manière volontaire ou non), où l'on associe fréquence élevée de certains codons avec un avantage sélectif. Sans imposer un quelconque avis sur le lien entre CUB et traduction, nous utilisons principalement ce terme lorsque celui-ci est employé par les auteurs cités.

1.2.1 *The genome hypothesis*

En 1980, *Grantham et al.* ont analysé le CUB d'ensembles de gènes appartenant à différentes espèces de virus, procaryotes, eucaryotes unicellulaires et multicellulaires animaux [32]. Les résultats obtenus les conduisirent à proposer une hypothèse appelée *the genome hypothesis* (l'hypothèse du génome), où le CUB est spécifique à chaque organisme et semble soumis à une quelconque sélection. Cette tendance se poursuivrait aux différents niveaux taxonomiques et notamment chez les mammifères où une universalité dans le contenu en GC des gènes serait, au sein de cette étude, démontrée [32]. Bien que ces résultats soient aujourd'hui discutables (comme nous le verrons par la suite, *H. sapiens* possède deux populations de CDS ayant des spécificités en CUB opposées), ils constituent l'un des premiers pas vers une analyse systématique du CUB avec l'idée d'y déterminer des tendances évolutives.

1.2.2 Sélection traductionnelle et disponibilité en ARNt

L'une des hypothèses les plus exploitées de sélection traductionnelle se base sur l'idée d'une association positive entre CUB, disponibilité des ARNt et fidélité et efficacité de la traduction [37, 46, 47]. Comme l'accès des différents ARNt au site A du ribosome est dominé par leurs quantités relatives, les ARNt disponibles en grande quantité ont une probabilité accrue d'être rapidement intégrés au sein du ribosome. Une certaine uniformité des codons synonymes afin qu'ils correspondent aux ARNt fréquents réduirait alors le temps d'attente de la complémentarité codon-anticodon, augmentant d'une part la rapidité de la traduction, mais aussi sa fidélité. Sur ce dernier point, la machinerie ribosomale peut être sujette à des erreurs d'attribution des acides aminés, et la probabilité qu'une erreur survienne augmenterait en fonction de l'attente de l'ARNt complémentaire du codon lu [48]. Il est à noter que selon l'organisme, les populations d'ARNt, et donc leur disponibilité au sein d'une cellule, sont classiquement déterminées par le nombre de copies de gènes associés : plus il y a de copies de gènes d'un ARNt, plus grande sera sa quantité dans le système cellulaire [49]. Une telle mesure est surtout vérifiée chez les procaryotes.

L'une des première mentions de l'importance des ARNt dans la modulation de l'expression des gènes fut émise par Itano en 1965, sans qu'aucune étude ne vienne appuyer cette hypothèse [50]. En 1975, *Fiers et al.* observèrent un lien entre CUB et expression des gènes chez le bactériophage (*i.e.* virus de bactérie) M52 d'*E. coli*. La protéine A et la protéine de la capside de ces virus, qui présentent une homologie dans la région 5' à l'échelle protéique mais pas à celle de l'usage des codons, sont effectivement exprimés à différents niveaux. Supposant une origine fonctionnelle à cette différence, *Fiers et al.* émirent l'hypothèse que le CUB influencerait l'expression des deux gènes. Néanmoins, ils ne firent aucune mention des populations d'ARNt dans leur étude [41]. Peu de temps après, *Efstratiadis et al.* firent état d'un usage non-aléatoire des codons de la b-globuline du lapin et établirent enfin un lien avec la disponibilité des différents ARNt iso-accepteurs [42]. Ils émirent tout de même des réserves à ce propos en soulignant la possibilité que les cellules eucaryotes soient capables de maintenir des populations d'ARNt tout au long de leur cycle cellulaire ; la notion de disponibilité des ARNt était donc pour eux désuette chez les eucaryotes [42]. Il faudra attendre les travaux de *Post et al.* pour voir apparaître un lien concret entre CUB, expression et abondance des ARNt. Ils proposèrent dans un premier temps

une hypothèse selon laquelle l'appauvrissement de certains ARNt isoaccepteurs au sein d'*E. coli* induirait une baisse de la rapidité et de la fidélité de la traduction de certains gènes [45, 46, 51]. En supposant que le contraire était tout aussi valable, ces mêmes auteurs démontrèrent la relation positive entre le CUB des gènes fortement exprimés des protéines *r* d'*E. coli* et la disponibilité des différents ARNt isoaccepteurs [46, 52].

Chez les procaryotes et les eucaryotes unicellulaires tels que *S. cerevisiae*, le lien entre CUB, abondance des ARNt et expression des gènes a été rapidement défini par Ikemura [31, 37] et Bennetzen et Hall [38]. En 1981, *Ikemura* analysa douze gènes d'*E. coli* pour lesquels l'expression était positivement corrélée à l'usage des codons et aux populations d'ARNt [37]. *Bennetzen et Hall* démontrèrent à leur tour l'existence d'un biais d'usage des codons chez deux gènes fortement exprimés de *S. cerevisiae*. Sur les 1004 acides aminés qui les composent, 96 % d'entre eux arborent un usage de 25 codons sur les 61 disponibles (ce qui implique un usage disproportionné de certains codons synonymes par rapport à d'autres). En les comparant à quatre autres gènes de levure moins exprimés, ils confirmèrent cette hypothèse et soulignèrent qu'un usage de codons associés à des ARNt rares sur des gènes fortement exprimés pourrait être délétère pour la cellule par épuisement des ARNt en question [38]. Dans un contexte plus général, de nombreuses études portant sur différents procaryotes ont amélioré nos connaissances sur le CUB de ces organismes. Une étude de *Sun et al.* présente une analyse de 61 génomes complets d'*E. coli* pour en déterminer le CUB des *core genes* (en français, les « gènes-coeur » à savoir les gènes partagés par tous les génomes d'un même organisme) et des gènes uniques à chaque génome [53]. Parmi toutes les explications valables quant aux différences de CUB entre les deux ensembles de gènes, *Sun et al.* proposent une origine évolutive : selon eux, les *core genes* ont un CUB orienté vers les abondances respectives des ARNt, ce qui n'est pas le cas des gènes plus spécifiques à chaque souche. Une telle différence pourrait s'expliquer par le fait que, au delà de leur expression accrue, l'ensemble des *core genes* contient des gènes essentiels à la survie d'*E. coli*. Leur expression doit donc être fidèlement assurée et exempte d'erreurs lors de la traduction [53]. Bien entendu, plusieurs autres facteurs peuvent expliquer une telle différence dans le CUB, telles que la différence de taille des séquences entre les deux catégories de gènes, ou encore des événements de transferts horizontaux [53]. La relation entre CUB et nombre d'ARNt jouerait par ailleurs un rôle prépondérant chez les bactéries ayant un court temps de génération [54]. Il a été en effet démontré par une étude sur les gènes de 102 espèces bactériennes que plus le temps de génération est court, plus le nombre de copies de gènes d'ARNt contenus dans le génome est grand, mais aussi que leur diversité est faible, de manière à ce que la traduction des gènes nécessaires à la croissance des individus (qui posséderaient alors un CUB en accord avec les populations d'ARNt) soit la plus optimisée possible [54]. Les résultats de cette étude vont jusqu'à trouver que les populations d'ARNt sont similaires chez les bactéries ayant un temps de génération court, et ce malgré leur distance phylogénétique. Ce dernier point suggère une universalité dans le CUB et dans l'abondance des ARNt au sein des procaryotes [54].

Chez les organismes eucaryotes multicellulaires, il est plus difficile de mettre en relation expression des gènes, usage des codons et abondance des ARNt dans la cellule. Il a été démontré que le CUB des gènes de l'araignée *Parasteatoda tepidariorum* varie en fonction de leur niveau

d'expression : les gènes fortement exprimés au sein de l'organisme possèdent un taux significativement plus élevé de T3 (T à la troisième paire de base du codon) que les autres gènes [39]. Parmi toutes les hypothèses possibles, *Whittle et Extavour* font état de la spécificité des populations d'ARNt pour expliquer une telle composition nucléotidique [39]. Chez la mouche du vinaigre *Drosophila melanogaster*, les mêmes conclusions ont pu être observées, où le CUB des gènes fortement exprimés est cette fois-ci orienté vers une présence accrue de codons enrichis en GC3 (G ou C à la troisième paire de base du codon) [55, 56]. Il est à noter que pour *Whittle et Extavour*, la différence de CUB entre *P. tepidariorum* et *D. melanogaster* pourrait être due à l'apparition de divergences dans les populations d'ARNt au cours de l'évolution des arthropodes [39].

Mais la complexité de ces organismes multicellulaires, et notamment le fait qu'ils arborent différents tissus cellulaires ayant tous une hétérogénéité qualitative et quantitative dans l'expression des gènes, nous force à revoir notre définition du CUB. Il doit être observé à l'échelle du tissu, et non pas de l'organisme entier. *Chevallier et Garel* observèrent en 1979 un changement de l'équilibre de certains ARNt isoaccepteurs à la fin du stade larvaire chez le bombyx du mûrier *Bombyx mori* [57]. Ce changement s'opère au niveau des tissus postérieurs et médians de leur glande séricigène, de manière à ce les populations d'ARNt se rapprochent du CUB des gènes de la fibroïne et de la séricine, soulignant ainsi une adaptation fonctionnelle et spatio-temporelle des ARNt au CUB de ces deux gènes fortement exprimés [57]. Sans entrer dans le débat entre hypothèses neutralistes et sélectionnistes, les auteurs de cette étude font état de la possibilité d'un effet mutationnel lié à la forte transcription de ces mêmes gènes (voir phénomène TAMB dans une section suivante) [57]. Pour aller plus loin dans cette démarche, *Whittle et al.* ont analysé le CUB de quatre tissus du tribolium rouge de la farine *Triboleum casteneum*, dont ceux de la lignée germinale [58]. Ils y ont découvert qu'en relation avec l'expression des gènes, les tissus de cet insecte possèdent des ensembles de codons « optimaux » quasi-identiques entre les tissus somatiques et de la lignée germinale, mais qui diffèrent lorsque l'on compare les deux types de tissus. De tels résultats pourraient signifier des différences de populations d'ARNt entre les tissus de *T. castaneum*, en accord avec l'expression tissu-spécifique des gènes [58].

Mais c'est chez les Vertébrés que le bât blesse : chez ces organismes, et en particulier les mammifères, le CUB est fortement impacté par le biais mutationnel [36, 43] (voir sections suivantes), et s'il existe un lien entre CUB, expression des gènes et disponibilité des ARNt, celui-ci est bien difficile à cerner. Quelques années après les travaux de *Chevallier et Garel*, *Hastings et Emerson* effectuèrent une analyse du CUB de deux ensemble de gènes « spécifiques », a priori, au foie et au muscle squelettique chez différentes espèces de Vertébrés, mais ne tirèrent pas les mêmes conclusions que leurs collègues : selon eux, il n'existe pas de différences significatives entre les gènes spécifiques à chacun de ces deux tissus au travers des Vertébrés, mais plutôt un consortium du CUB au sein des gènes fortement exprimés, marqué par un enrichissement en GC3 propre à la *genome hypothesis* de *Grantham et al.* [32, 59]. La majeure partie des études qui ont suivi n'a jamais démontré formellement l'existence d'une sélection traductionnelle des codons synonymes chez les Vertébrés, qui plus est en corrélation avec l'abondance des ARNt. C'est le cas d'une étude récente de *Pouyet et al.* qui, au détour d'une analyse de l'impact du gBGC (*GC-biased gene conversion* en anglais ou biais de conversion génique GC-biaisé en fran-

çais, discuté à la suite de ce rapport) sur le CUB des gènes humains précise l'absence d'un lien entre abondance des ARNt et CUB [43]. Cela dit, certains auteurs tentent toujours d'élucider la question de l'existence d'une sélection traductionnelle chez les Vertébrés : pour *Yi et al.* le taux de GC3 et l'usage des codons de trois espèces de loches sont étroitement liés à l'expression des gènes fortement exprimés, mais toujours sans que cela ne soit mis en rapport avec l'abondance des ARNt [60]. Une hypothèse indiquée dans une étude de *Duret et Mouchiroud* propose que l'absence de signes de sélection traductionnelle chez *H. sapiens* pourrait être partiellement expliquée par les différences de taille de population avec d'autres organismes tels que *D. melanogaster*, *S. cerevisiae* ou encore *E. coli* : une mutation avantageuse pour une espèce avec une grande population pourrait être *de facto* neutre au sein d'une population plus petite, où la dérive génétique l'emporte sur la sélection, spécialement si le coefficient de sélection est petit [44].

Comme nous le verrons par la suite, les organismes cellulaires, dont les mammifères, peuvent être soumis à des fluctuations dans l'abondance de leurs ARNt au cours du cycle cellulaires ou sous certaines conditions [61]. De telles observations pourraient alors remettre en question notre interprétation du lien entre CUB, abondance des ARNt et expression des gènes, où la relation entre ces trois variables pourrait être profondément modifiée en fonction de la condition de la cellule.

1.2.3 Effet du CUB sur la stabilité et la maturation des ARNm

Il n'est pas possible d'étudier l'impact du CUB sans évoquer son influence sur la composition nucléotidique des ARNm. Pendant et après la transcription, la composition globale et locale des ARNm induit leur structure et leur maintien, ce qui participe à la régulation de l'expression des gènes [62]. La stabilité d'un ARNm (*i.e.* sa demi-vie) dépend largement de sa structure secondaire, elle même définie par les possibles interactions de l'ARNm avec lui même.

Chez la levure *S. cerevisiae*, l'insertion de codons synonymes rares au sein du gène rapporteur *PGK1* peut conduire à une baisse du maintien de l'ARNm correspondant : le ralentissement de l'étape d'élongation induit par le CUB peut favoriser, *via* l'action de la protéine inhibitrice de la traduction Dhh1, le *decapping* (décoiffement) de l'ARNm, menant alors à sa dégradation [63]. Un tel phénomène serait associé à la vitesse de la traduction et donc à l'activité des ribosomes en contact avec l'ARNm : si celui-ci possède un nombre important de codons rares, les ribosomes s'accumulent et se bloquent lors de l'étape d'élongation, provoquant alors un *ribosome-jam* (embouteillage de ribosomes) par *ribosome stalling* (arrêt du ribosome) (Figure 1.4). L'action de la protéine Dhh1 permettrait alors la libération des ribosomes, de manière à ce que la traduction des autres gènes ne soit pas impactée par le *ribosome-jam* [63]. Mais le contraire est tout aussi valable : lors d'une analyse sur des versions « optimisées » et « désoptimisées » (*i.e* modifiées pour présenter des codons synonymes fréquents ou rares) du gène *HIS3* du même organisme, il a été remarqué que la demi-vie des ARNm était drastiquement augmentée chez les transcrits « optimisés », et ce par le biais d'une augmentation de la vitesse de translocation lors de la phase d'élongation [34]. En comparaison avec la version *wild-type* (naturelle) du gène *HIS3*, la version « optimisée » de *Presnyak et al.* a une demi-vie près de six fois supérieure, alors que la version « désoptimisée » en a une cinq fois moindre. Pour chacun de ces deux exemples, il semblerait que le biais d'usage des codons agisse, par le biais de la modulation de la vitesse de la traduc-

tion, comme un régulateur de la stabilité d'un ARNm via une étroite relation entre activité des ribosomes et machinerie de dégradation des ARNm [34].

Selon la structure secondaire d'un ARNm, les étapes d'initiation et d'élongation de la traduction peuvent être modifiées. Par exemple, la région d'initiation de la traduction RBS (*Ribosome Binding Site* ou site de Fixation du Ribosome en français) du site d'initiation de Shine-Delgarno d'*E. coli* peut être inaccessible en fonction de sa composition nucléotidique, au point d'empêcher toute traduction du gène [64, 65]. En effet, le contenu en AUGC d'un ARNm peut conduire à la formation de *hairpins* qui peuvent bloquer l'accès du site de Shine-Delgarno aux ribosomes [65]. Les mêmes conclusions peuvent être tirées quant à l'impact des *hairpins* sur l'initiation de la traduction chez les eucaryotes : la formation de structures particulières dans la région 5' d'un ARNm peut conduire à une modulation de l'expression des gènes, comme l'indiquent *Tuller et Zur* dans leur revue [66]. Il est toutefois indiqué dans ce même article que certains gènes de *S. cerevisiae* possèdent après le site d'initiation des *hairpins* avec une forte cohésion. On pourrait penser qu'une telle structure serait délétère pour la traduction du gène, mais il semblerait que cette région favoriserait au contraire l'initiation de la traduction en i) empêchant la formation de *hairpins* autour du site d'initiation ; ii) favorisant le trafic des ribosomes le long de l'ARNm [66]. Chez les eucaryotes, le site d'inititation est reconnu par la sous-unité 40S du ribosome. Cette sous-unité possède une affinité avec une région dite de Kozak qui permet l'identification du site d'initiation [66, 67]. Selon la composition nucléotidique de la région de Kozak, le site d'initiation peut être plus ou moins bien reconnu par le ribosome, auquel cas *Kozak* proposa deux grandes explications : i) le ribosome se décale et trouve un autre codon initiateur aux abords du premier et véritable site d'initiation ; ii) le ribosome se détache de l'ARNm et ne synthétise pas la protéine [67]. La région de Kozak est très fortement conservée, et ses modifications, bien que rares, peuvent provoquer des afflictions chez *H. sapiens* : la mutation d'un simple nucléotide dans une position conservée de cette séquence peut amener à une modification de la protéine *SOX9* par sélection d'un autre codon initiateur que celui d'origine, ce qui peut entraîner des cas de dysplasie campomélique [68].

Chez les eucaryotes, où les ARNm sont soumis à une étape de maturation comprenant un épissage (*i.e* une coupure de l'ARNm pour en supprimer les introns), l'usage des codons à la frontière entre les exons et les introns jouerait un rôle dans ladite maturation de l'ARNm. D'un point de vue général, ces régions sont composées de codons AT3-riches associés à des contraintes sur les limites des exons [69, 70]. Par exemple, le CUB retrouvé aux abords des exons de la protéine *TP53* aide à son épissage par le concert des protéines régulatrices du phénomène [71]. Un bouleversement du CUB dans ces régions pourrait alors empêcher la formation d'ARNm matures par absence d'un épissage alternatif correct.

1.2.4 Variations intragéniques du CUB

Comme nous avons pu le voir précédemment, les variations dans le CUB peuvent, selon l'organisme, influencer la vitesse et la fidélité de la traduction des ARNm [31, 38, 54]. Il existe aussi des variations intragéniques qui peuvent moduler le rythme d'élongation de la traduction tout au long de l'ARNm. Chez certains gènes fortement exprimés de procaryotes et d'eucaryotes, la région suivant le site d'initiation est composée de codons rares (Figure 1.4). L'interprétation la

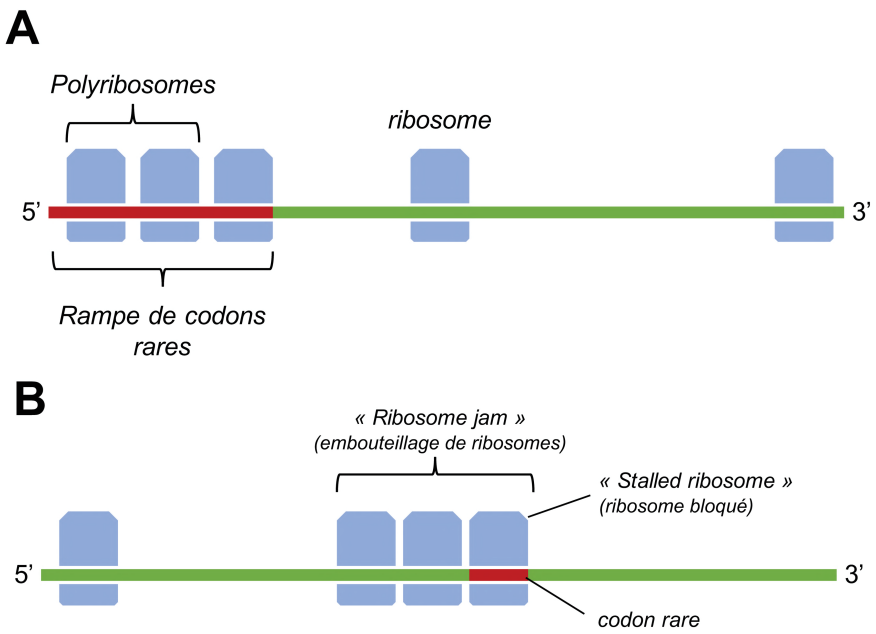

Figure 1.4 – **Représentation schématique A) de la « rampe » de codons rares et B) du ribosome-jam.** A) La rampe de codons rares se situe dans la région 5' de l'ARNm (région dans laquelle les ribosomes s'assemblent autour de l'ARNm). De par la rareté des codons et des ARNt associés dans cette région, la rampe freine le rythme de la phase d'élongation de la protéine pour empêcher une accumulation trop importante de ribosomes. Une fois la rampe dépassée, les ribosomes gagnent en vitesse et effectuent une élongation plus rapide. B) Le *ribosome-jam* naît de l'arrêt soudain (*ribosome stalling*) ou de la baisse de régime d'un ribosome au cours de la phase d'élongation de la traduction. Le ralentissement du ribosome provoque une accumulation de ribosomes, une baisse générale du rythme d'élongation de l'ARNm et une pénurie de ribosomes au sein de la cellule.

plus habituelle de cette particularité dans le CUB est que cette région, appelée « rampe » et dont la taille varie entre 30 et 50 codons, empêcherait l'apparition d'un *ribosome jam* et permettrait donc une traduction optimale des gènes (Figure 1.4) [72]. Après cette région « rampe », le CUB du gène s'« optimise » à nouveau pour regagner un usage des codons propre aux gènes fortement exprimés. Il faudrait donc voir la « rampe » de CUB comme une voie d'insertion d'autoroute où les ribosomes démarrent lentement l'élongation de la protéine puis acquièrent une vitesse de pointe où les polyribosomes (*i.e.* la succession de plusieurs ribosomes sur un même ARNm) ne se gêneront pas [2, 72].

Chez la levure *S. cerevisiae*, les protéines secrétées par la cellule ainsi que celles de la membrane plasmique possèdent des groupements de codons rares à une distance de 30 à 45 codons par rapport aux sites de liaison de la SRP (*Signal Recognition Particle* en anglais ou particule de reconnaissance du signal en français) [2] [73]. La SRP reconnaît les signaux de sécrétion des pro-

téines, ou encore les segments transmembranaires des protéines de la membrane dès leur sortie du ribosome. Une fois reconnues par les SRP, les protéines naissantes sont acheminées vers les complexes de translocation des protéines afin d'être déplacées au sein de la cellule. Le groupe de codons rares, qui est à une distance équivalente à la longueur du tunnel de sortie des ribosomes (*i.e* de la partie intra-ribosomale de la chaîne polypeptidique), aurait été sélectionné pour laisser l'espace et le temps nécessaires à la reconnaissance du site de liaison de la protéine par le SRP, et ce via un ralentissement de l'élongation de la protéine [73].

1.2.5 Variations intergéniques du CUB

La taille d'une séquence pourrait être un facteur explicatif du CUB. Contrairement aux autres exemples donnés ci-dessus, où la rapidité de la traduction est un facteur prédominant dans la sélection des codons, le postulat de départ est qu'un CUB « optimisé » sur de longues séquences relèverait d'une stratégie d'économie d'énergie de la cellule [74]. En effet, plus une séquence est longue, plus l'étape d'élongation peut induire des erreurs de traduction délétères pour la protéine naissante – comme l'insertion d'acides aminés imprévus ou l'arrêt précoce de la traduction – et donc coûteuses pour la cellule de par l'investissement énergétique dans une protéine non-fonctionnelle. Pour éviter cela, les grandes séquences auraient généralement un CUB riche en codons fréquents. Une étude portant sur les gènes de *S. cerevisiae*, *D. melanogaster* et *E. coli* a mis en évidence une corrélation positive et significative entre l'utilisation de codons fréquents et la longueur des gènes chez la bactérie, mais pas chez les deux eucaryotes [74]. Chez les procaryotes, il est donc possible que l'hypothèse émise en début de paragraphe soit valide. Une étude de 1999 de *Duret et Mouchiroud* confirme ces résultats, où les longs gènes d'*Arabidopsis thaliana*, *Caenorhabditis elegans* et *D. melanogaster* voient leur « optimalité » baisser par rapport aux gènes courts et ce pour des niveaux d'expression sensiblement similaires [44]. Plusieurs hypothèses de sélection traductionnelle ont été émises et vérifiées au cours de cette étude, sans qu'aucune n'explique véritablement les observations faites sur ces organismes. *Moriyama et Powell* proposent tout de même une hypothèse où la taille des séquences est contre-sélectionnée chez les gènes fortement exprimés [74]. Ici, *Moriyama et Powell* supposent que le mécanisme d'évitement des erreurs lors de la traduction se fait indirectement par une sélection de protéines ayant la même fonction mais dont la taille serait réduite [74]. Mais de tels résultats ne sont pas présents au sein de l'étude de *Duret et Mouchiroud*, car la taille des séquences ne semble pas corrélée à l'expression des gènes des trois organismes étudiés [44]. Au final, il semblerait que les gènes bactériens soient soumis à une pression de sélection positive des gènes « optimaux » en fonction de la longueur de la séquence, alors que plusieurs facteurs semblent provoquer une tendance inverse chez les eucaryotes [44, 74].

À l'instar des ARNm et des protéines, les populations d'ARNt sont soumises aux fluctuations temporelles, spatiales et environnementales d'une cellule. Au cours du cycle cellulaire, d'un stress ou de conditions spécifiques, ces populations peuvent être notablement modifiées et changer ainsi l'expression des gènes [61]. La synthèse des ARNt et leur circulation dans la cellule implique un grand nombre d'étapes de maturation et de trafic cellulaire pouvant toutes être modulées selon le stress induit. En suivant l'hypothèse de sélection traductionnelle, les dif-

férentes populations de gènes disposant d'un CUB différentiel peuvent s'optimiser en fonction de la population des ARNt disponibles. Une étude de *Torrent et al.* s'est focalisée sur l'analyse des populations d'ARNt, d'ARNm et de l'expression des gènes sous trois conditions de stress chez la levure du boulanger. Les gènes fortement exprimés à la suite d'un stress possèdent, dans des conditions normales, des codons associés à des ARNt rares. Sous l'effet d'un stress, les populations d'ARNt varient chez la levure, et les codons rares deviennent « optimisés », ce qui améliore la traduction des gènes de réponse au stress [75]. Chez *H. sapiens*, les gènes exprimés aux différentes étapes du cycle cellulaire ont un CUB riche en codons rares [76]. Pour être plus précis, ces gènes possèdent par exemple une forte proportion du codon UUU-Phe (qui n'a pas d'ARNt associé chez l'Homme). Or, ces codons peuvent être considérés comme « rares », car l'appariement avec un ARNt *near-cognate* ne se fait pas aussi efficacement qu'avec un ARNt *cognate*. Mais dès que la quantité d'ARNt augmente (ce qui est le cas pendant la phase G2 du cycle cellulaire), les codons UUU-Phe sont décodés avec autant d'efficacité que les codons UUC-Phe, car la saturation des ARNt associés permet un meilleur appariement entre codons et ARNt *near-cognate* [76]. De ce fait, les codons *wobble* s'« optimisent » selon la phase du cycle cellulaire. Il est à noter que les gènes exprimés pendant la phase G1 (de croissance cellulaire) sont « optimisés » sur tout le long de leur séquence. Chez ces gènes, il n'existe pas de « rampe » de codons. L'explication la plus satisfaisante est que la population des tous les ARNt est au plus bas lors de cette phase, et qu'il est absolument nécessaire d'exprimer ces gènes [76]. Bien entendu, il est primordial de considérer une sélection traductionnelle chez *H. sapiens* pour accepter de telles hypothèses.

1.2.6 Relations inter-codons

La co-occurence des codons se définit par la répétition de codons synonymes dans une courte région (Figure 1.5 A). Lors de la traduction, les ARNt ayant récemment transféré leur acide aminé à la protéine en formation restent à proximité de la machinerie ribosomale. Dans certains cas, ils peuvent rapidement être munis d'un nouvel acide aminé et, si le ribosome rencontre un nouveau codon correspondant, être directement réutilisés par celui-ci [2, 77]. De ce fait, la co-occurence des codons représente virtuellement une population locale d'ARNt pouvant être recyclés pour améliorer la traduction de manière ponctuelle. Pour démontrer cela, *Cannarozzi et al.* ont effectué une analyse de la co-occurrence des codons chez *S. cerevisiae* [77]. En comparant l'expression de deux gènes modifiés de la GFP (l'un avec une forte co-occurrence, l'autre sans), ils ont remarqué l'importance de la proximité de codons identiques dans la vitesse de traduction, et ce même pour des codons associés à des ARNt rares [77].

Chez les Vertébrés, les fréquences différentielles des paires de codons (*i.e.* deux codons consécutifs), ou dicodons, à savoir NNN-NNN, ne sont pas du ressort *stricto sensu* du CUB mais peuvent l'influencer en orientant le choix des codons synonymes qui les composent (Figure 1.5 B). Ce phénomène, appelé biais de paires de codons, provient d'un déséquilibre dans la fréquence des paires de codons par rapport à d'autres : chez *H. sapiens*, la paire de codons GCC-GAA est très peu utilisée par rapport aux 12 autres combinaisons possibles Alanine-Glutamate, malgré le

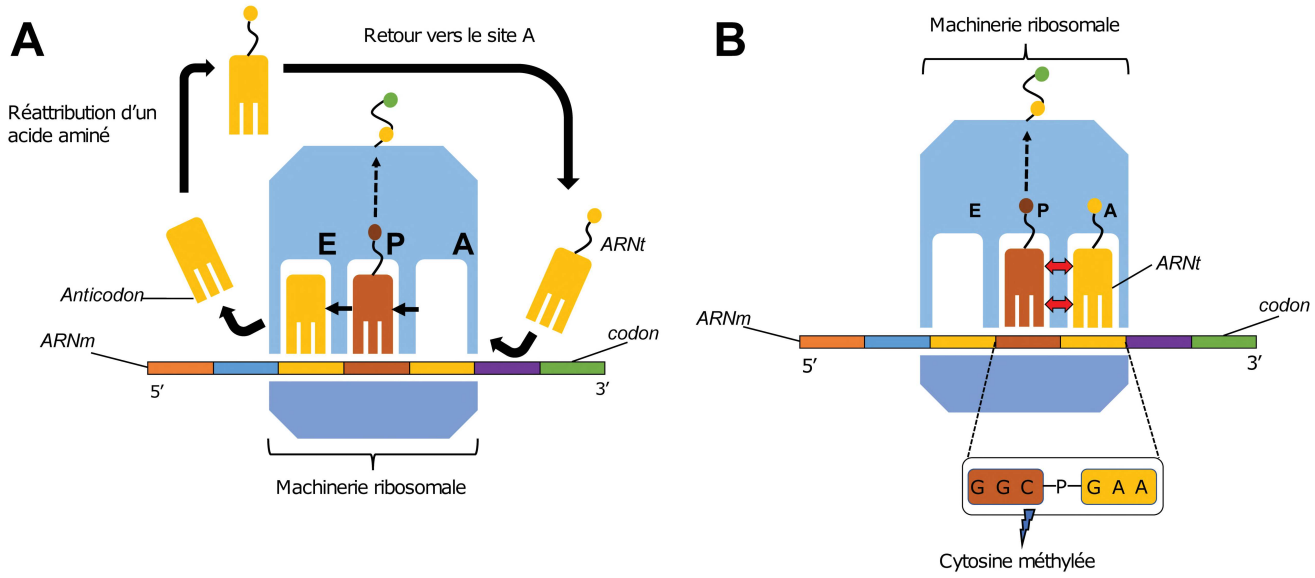

Figure 1.5 – **Relations entre les codons au sein d'une même séquence.** A) Le biais de co-occurence des codons résulte de la répétition d'un même codon synonyme sur une courte séquence de l'ARNm. Après sélection de l'acide aminé de l'ARNt qui s'est associé à la première occurrence du codon synonyme, celui-ci peut gagner rapidement un nouvel acide aminé et être rappelé par la machinerie ribosomale lors de la lecture d'une nouvelle occurrence du codon synonyme. B) Le biais d'usage des paires de codons implique l'existence de dinucléotides retrouvés à la frontière entre deux codons. Selon leur nature, les dinucléotides peuvent être contre-sélectionnés par des mécanismes de dégradation et de réparation de l'ADN (comme les CpG méthylés et la désamination de la cytosine du dinucléotide en thymine). De ce fait, il existe aussi un biais d'usage des paires de codons, où certains sont contre-sélectionnés par rapport à d'autres. Bien entendu, un tel aspect influence le CUB de la séquence.

fait que le codon GCC soit le plus utilisé pour coder l'alanine au sein du génome humain [78]. Plusieurs hypothèses expliquent un tel déséquilibre dans l'usage des paires de codons.

Une première suppose que lors de l'élongation de la protéine, l'interaction entre les ARNt contenus dans les sites A et P peut avoir une influence sur le déroulement de la traduction [2, 79]. Par ailleurs, chez les procaryotes, les paires composées de codons rares sont sur-représentées alors que celles de codons optimisés sont rares. Les hypothèses expliquant un tel inversement de tendance sont multiples :

i Dans certains cas, l'association de deux codons rares favoriserait un ralentissement de la traduction favorable à un repliement correct de la protéine [2, 79].

ii Les paires de codons rares peuvent participer à la mise en place d'un système d'arrêt de la traduction et de libération des ribosomes par dégradation de l'ARNm via l'activité de la ribonucléase RelE et de l'ARNtm SsrA (ARN de transfert-messager), et ce sous certaines

conditions [79].

Chez *E. coli*, l'activité des ARNtm est stimulée, par exemple, par la présence d'une paire de codons AGA-AGA, où AGA est un codon rare pour l'arginine [79]. L'ARNtm, en marquant cette paire de codons rares, induirait la formation d'une protéine incomplète et l'arrêt de la traduction [79, 80]. Mais il est tout à fait possible que ce mécanisme ne soit pas forcément lié au couple AGA-AGA, mais plutôt à la répétition d'un codon rare qui conduirait à l'incorporation d'un autre acide aminé que l'arginine et provoquerait alors la formation d'une protéine non-fonctionnelle [48]. Dans le cas où la quantité d'ARNt AGA est appauvrie, le même système est appelé pour un seul et unique codon. Dans ce cas, il est supposé que ce mécanisme protégerait d'un *ribosome-stalling*. C'est donc l'accumulation de codons rares qui serait à l'origine d'un tel phénomène plutôt que la formation d'une paire de codons [80]. On pourrait alors suggérer qu'en cas de disette des ARNt associés à cette paire de codons, le mécanisme d'arrêt de la traduction serait en réalité un système de protection cellulaire visant à libérer les ribosomes, et ce pour permettre une meilleure traduction des autres gènes [80].

Mais le CUB induit par les paires de codons pourrait tout aussi bien exister via les déséquilibres des dinucléotides à la frontière des deux codons. Parmi toutes les combinaisons possibles, les CpG (à savoir Cytosine-phosphate-Guanine) sont considérés comme « hyper-mutables » (*i.e* sujets à un fort taux de mutation) car potentiellement soumis à une méthylation de la cytosine qui peut, *in fine*, conduire à la désamination du nucléotide méthylé en une thymine (et donc à l'apparition d'un dinucléotide TpG) [81]. La méthylation des nucléotides est un processus épigénétique que l'on retrouve chez les procaryotes et les eucaryotes et qui peut conduire à des modifications de l'expression des gènes : chez les eucaryotes, un gène possédant une faible méthylation est transcrit normalement alors qu'un gène fortement méthylé, notamment au niveau de ses promoteurs, voit sa transcription inhibée [82]. Chez les Vertébrés, et à l'exception des îlots de CpG, ces dinucléotides sont sous-représentés dans l'ensemble du génome et sont pour cela en fréquence réduite à l'échelle des paires de codons. Mais chez d'autres organismes eucaryotes, comme *D. melanogaster*, l'évitement des CpG est, en comparaison avec les Vertébrés, amoindri. Cela peut notamment s'expliquer par l'absence de certaines enzymes de la famille des *DNA methyltransferases* (ADN méthyltransférases en français) qui sont responsables de la méthylation des CpG [83]. Il est à noter que de nombreux virus miment leurs hôtes en évitant eux aussi les combinaisons CpG, et ce pour permettre une meilleure expression de leurs gènes mais aussi pour éviter des mécanismes de défense des hôtes [2, 78, 84]. En effet, la protéine ZAP (pour *Zinc-finger Antiviral Protein* ; protéine antivirale en doigt de zinc en français) viserait les sites CpG des virus à ARN et inhiberait l'expression des gènes et la réplication virale par de multiples mécanismes [85]. Certains auteurs suggèrent même de modifier le CUB des gènes de certains virus vers un enrichissement en CpG en guise de solution vaccinale, car les gènes viraux ne seraient que peu exprimés et leur virulence en serait atténuée [78, 85].

1.2.7 Effet du CUB sur le repliement des protéines

Le repliement d'une protéine est le processus durant lequel elle acquiert une structure tridimensionnelle lui conférant sa fonction. Les relations biochimiques entre les différents acides aminés (hydrophilie, hydrophobicité, charge électrique) sont tout autant d'éléments qui vont

conduire à la formation de structures secondaires et tertiaires de la protéine. Les protéines chaperonnes, comme GRoEl, participent au repliement des protéines et aident donc à l'acquisition de leur fonction.

La distribution des codons synonymes au sein des gènes procaryotes peut permettre, selon les cas, de modifier le rythme de l'élongation de la protéine, ce qui peut jouer un rôle dans la formation de leur structure [69, 86, 87]. Les gènes dont les protéines possèdent au sein de leur structure tridimensionnelle des hélices α décrivent un CUB particulier : au début de l'hélice, les codons aux positions 1 et 4 sont systématiquement rares alors que les codons 2 et 3 sont fréquents. Une telle alternance pourrait souligner les spécificités du repliement de l'hélice α qui se déroule à la sortie du ribosome pendant l'élongation. Les feuillets β sont quant à eux enrichis en codons fréquents et ne démontrent pas de particularités entre codons rares et fréquents [88]. Chez *Echinococcus granulosus*, la protéine EgFABP possède une succession de deux hélices α . En transfectant différentes versions synonymes de ce gène chez *E. coli*, *Cortazzo et al.* remarquèrent que la protéine a une probabilité accrue de mal se replier si les codons rares entre les deux hélices α sont remplacés par des codons plus fréquents [86]. Chez l'eucaryote *S. cerevisiae*, la modification des codons synonymes rares de l'enzyme TRP3 conduit à une baisse de son activité enzymatique, probablement par le biais d'une acquisition de structures secondaires et tertiaires bancales [87]. Un changement de l'équilibre entre codons rares et fréquents à des positions cruciales pour le repliement des protéines pourrait donc avoir un effet délétère sur le gain de fonction des protéines en bouleversant le rythme de l'élongation [86, 87]. Pour aller encore plus loin, nous pouvons citer l'exemple de *Shah et al.* sur le gène *CFTR*, qui est partiellement responsable de la muscoviscidose lorsqu'incorrectement exprimé [89]. Celui-ci semble voir son repliement impacté lorsque le CUB est « optimisé » pour qu'il soit composé de codons plus fréquents au sein de l'organisme [89]. Une des hypothèses soumise par *Shah et al.* est qu'en l'absence de codons rares, l'étape d'élongation est plus rapide et ne laisse pas le temps aux protéines chaperonnes de replier correctement certaines régions de la protéine en formation [89]. Encore une fois, de telles suppositions admettent en postulat de départ qu'il existe une sélection traductionnelle chez *H. sapiens*.

1.2.8 Biais mutationnel et gBGC

Le biais mutationnel correspond à un déséquilibre des mutations par rapport à une attente aléatoire. Au sein de cette sous-section, nous décrirons une liste exhaustive de biais mutationnel et les placerons dans le contexte du CUB.

La mesure des forces mutationnelles est majoritairement réalisée à l'échelle des SNP (*Single Nucleotide Polymophism* en anglais ou polymorphisme nucléotidique simple en français) des populations d'une même espèce. Différentes études de génétique des populations ont démontré qu'un polymorphisme observé à l'échelle des populations d'une même espèce ou d'espèces proches est beaucoup moins représentatif d'une pression de sélection qu'entre individus d'espèces éloignées et est moins soumis à des événements façonnant la composition nucléotidique tels que le GBC (*Gene Bias Conversion*, voir ci-dessous). Dans ce cas, la mesure du polymorphisme est donc un excellent estimateur du biais mutationnel d'une espèce ou d'une population [90–92]. L'analyse inter-générationnelle des SNP permet de connaître les différents taux de muta-

tions, mais aussi leur nature au sein d'une espèce. Or, il existe chez les eucaryotes et procaryotes un déséquilibre universel des mutations vers un enrichissement en AT [92, 93], et ce malgré une forte diversité dans le contenu en GC. À titre d'exemple, les organismes *Streptomyces coelicolor* et *Plasmodium falciparum* arborent tous deux un CUB orienté vers un usage quasi-strict de codons se terminant en GC pour le premier et en AT pour le deuxième [94]. En analysant l'accumulation des mutations sur des génomes complets de lignées appartenant à 37 espèces eucaryotes et procaryotes, *Long et al.* confirment l'aspect universel de ce déséquilibre, et décrivent une pseudo-universalité des mécanismes, notamment de sélection et de biais mutationnels, favorisant ainsi la fixation de régions enrichies en GC dans certains génomes [95]. Ainsi, le gBGC, la méthylation des dinucléotides CpG (voir ci-dessous) et la taille de population efficace sont tout autant d'origines pouvant expliquer la diversité en contenu en GC au sein de l'arbre du vivant [95]. Une analyse se focalisant cette fois-ci sur la diversité de cinq espèces de bactéries pathogènes ayant un mode de reproduction clonal démontre une accumulation des mutations GC -> AT au sein de chaque espèce, et ce malgré une diversité dans leur contenu en GC [92]. Comme pour *Long et al.*, *Hershberg et Petrov* supposent que c'est la modulation du rythme des mutations ainsi que l'impact d'autres événements façonnant la composition nucléotidique qui déterminent le contenu en GC d'un génome bactérien et de son CUB : les génomes de ces cinq espèces bactériennes tendent tous vers une accumulation des substitutions provoquant un enrichissement en AT, mais celle-ci est contrebalancée par différents événements qui modifient la composition nucléotidique [92]. Le choix de ces organismes pour une telle étude s'explique par le fait que :

i certains d'entre eux, comme les souches MTBC (*Mycobacterium tuberculosis cluster*), possèdent une forte proximité phylogénétique, constituant un parfait terreau pour une analyse sur les forces mutationnelles chez les bactéries,

ii le mode de vie et la reproduction clonale de ces espèces semble alléger les contraintes sélectives sur ces espèces.

Les auteurs de cette étude expriment néanmoins quelques réserves au sujet de leur étude, où certains gènes restent tout de même fortement conservés. Pour finir, ces auteurs indiquent que la désamination de la cytosine méthylée (voir ci dessous) pourrait être l'une des raisons majeures de l'universalité des transitions GC vers AT [92].

Chez les eucaryotes, les mêmes conclusions peuvent être déduites à partir de plusieurs études. Chez la drosophile *D. melanogaster*, *Haddrill et Charlesworth* tirent les mêmes conclusions que *Hershberg et Petrov* : dans les régions non-codantes où le gBGC et la sélection naturelle ont un impact limité sur la composition nucléotidique, les mutations favorisent l'installation de nucléotides AT. Au contraire, les autres régions non-codantes arborent un contenu en GC bien plus élevé, et ce probablement car la pression de sélection et les recombinaisons conduisent à une augmentation du contenu en GC chez cette espèce [93, 96]. Ces résultats sont confirmés par l'analyse du rétrotransposon *Helena* présent dans les génomes de *D. melanogaster* et *Drosophila virilis*, où malgré une divergence phylogénétique marquée, chaque espèce étudiée possède un taux de transition GC->AT plus élevé que le reste des substitutions au sein du rétrotransposon [97]. Il est à noter que le rétrotransposon *Helena* est un élément génétique *dead-on-arrival* (« mort à l'arrivée » en français) qui agit comme un pseudo-gène et n'est donc pas soumis à une quelconque pression de sélection [97]. La même étude confirme l'existence d'un biais mutationnel

similaire mais bien plus élevé chez les mammifères, et détermine que cette différence dans le taux de transitions GC vers AT est majoritairement liée à la désamination des cytosine méthylées (voir ci-dessous). De manière intéressante, cette étude démontre une équiprobabilité dans les taux des autres substitutions, suggérant de fait une universalité dans les taux de mutations chez les bilatériens [97].

La réplication de l'ADN d'un organisme cellulaire s'effectue sur les deux brins dits directs et indirects via l'activité des ADN polymérases [98]. Ce processus est effectué à l'aide de facteurs de réplication qui vont notamment provoquer l'ouverture de l'ADN en une « fourche de réplication » pour permettre la synthèse des deux nouveaux brins. La réplication du brin direct se fait dans le sens de l'ouverture de la fourche de réplication, et suit donc les mécanismes qui y sont associés. En contre-partie, la synthèse du brin indirect se fait de manière morcelée par la création de fragments dits d'Okazaki. Ces fragments, qui sont détachés les uns des autres, sont regroupés pour former une seule macromolécule d'ADN par le biais de ligases. Malgré les capacités de relecture des ADN polymérases, les deux brins directs et indirects fraîchement synthétisés sont sujets à des erreurs lors de la réplication qui représentent une forte majorité des mutations de la cellule, notamment au niveau des sites de liaison des fragments d'*Okazaki*. Ces erreurs sont détectées et corrigées par le système protéique MR (pour *mismatch-repair* ou réparation de mésappariements en français) sur le brin en cours de synthèse avec comme modèle le brin parent [99]. Mais, à une certaine mesure, la réparation de l'ADN peut mener à la modification de l'ADN, provoquant alors un biais mutationnel à l'échelle du génome entier.

Le taux de transcription des gènes peut influencer leur CUB. Lors de la transcription, le brin transcrit et le brin non-transcrit de la molécule d'ADN se séparent. Le brin non-transcrit est alors temporairement monocaténaire et des épisodes de désaminations de cytosines non-méthylées en uraciles sont provoqués par les désaminases cellulaires. Les uraciles issues de cette désamination sont par la suite remplacées par des thymines, de manière à conserver la structure de l'ADN [100]. Ce phénomène, qui conduit obligatoirement à l'apparition d'un mésappariement entre les nucleotides G-T des deux brins, est la plupart du temps corrigé par un mécanisme de réparation post-transcriptomique pour que l'ADN retrouve sa séquence initiale. Mais cette correction peut conduire au remplacement du nucléotide G par un A, ce qui encourage un biais mutationnel GC->AT par transcription des gènes. Ce phénomène est appelé TAMB (pour *Transcription-Associated Mutational Biases* ou biais mutationel associé à la transcription en français), et prendrait principalement place dans les gènes exprimés au sein des tissus de la lignée germinale, bien qu'il soit observé à une moindre mesure dans les gènes exprimés au sein d'autres tissus [100, 101]. On peut estimer que plus un gène est transcrit, plus il est soumis au TAMB et sera donc enrichi en AT [39].

Les dinucléotides CpG, précédemment indiqués comme contre-sélectionnés pour permettre une meilleure expression des gènes, sont aussi soumis au phénomène de désamination de la cytosine [102]. Lorsqu'elle est méthylée, la cytosine du dinucléotide peut être remplacée par une thymine via l'action des désaminases [102]. Bien que ce phénomène ne soit pas en contradiction avec une quelconque pression de sélection sur le dinucléotide (et voire même y participerait), son

existence façonne le CUB des gènes en modifiant le taux global de GC observé et demeure l'une des mutations les plus courantes au sein des organismes, favorisant alors les transitions GC->AT.

Chez les eucaryotes, et à l'instar des biais mutationnels, le gBGC est un mécanisme à l'origine d'importants changements dans la composition nucléotidique observée. Ce processus, qui se déroule lors d'une recombinaison (somatique ou meïotique), induit une modification positive du contenu en GC d'un gène. Il correspond en une modification de la probabilité de fixation des allèles, généralement positive pour les GC-riches, et ce via le phénomène dit de conversion génique (*i.e* transfert d'information génétique non-réciproque) [103]. Cette conversion prend place lors de la réparation double-brin d'une région d'ADN endommagée, et elle correspond en la copie de la région équivalente d'une autre chromatide (Figure 1.6). Le fragment endommagé est tout d'abord digéré sur ses deux brins dans le sens 3'-5', de manière à ce que les régions entourant la cassure soient simple-brins. Grâce à la structure monocaténaire du fragment, celui-ci forme des hétéroduplexes avec la région de la chromatide qui lui est complémentaire. La réparation des brins d'ADN est effectuée par synthèse des régions simple-brins avec comme modèle la séquence de la chromatide apparentée. Le complexe obtenu est appelé jonction d'Holliday et peut, dans certains cas, subir des événements de *crossing-over* (enjambement ou entrecroisement en français) [103]. Dans le cas où les régions endommagées et non-endommagées différaient à la base, le mécanisme de réparation de la cassure double-brin engendre ladite conversion de l'ADN endommagé, car cette région devient une stricte copie de son complémentaire. Deux mécanismes semblent être majoritairement à l'origine de la conversion des gènes par réparation : le système NER (Nucleotide Excision Repair) qui digère puis resynthétise fidèlement l'ADN endommagé à partir de l'ADN non-endommagé et le système BER (Base Excision Repair) qui répare les inadéquations entre les deux brins des hétéroduplexes [103]. C'est ce dernier système qui serait grandement responsable du gBGC, car il favoriserait le remplacement des allèles riches en AT par des allèles riches en GC lors de la réparation de l'ADN, que ce soit sur le brin endommagé ou le brin modèle. Il a été démontré que le système BER possède un rôle majeur dans le gBGC observé chez la souris commune *Mus musculus* [104]. Mais chez *S. cerevisiae*, ce même système ne semble pas être corrélé aux patrons de gBGC observés [104, 105]. Par ailleurs, lors de la recombinaison meïotique, les allèles contenant des AT sont plus sujets à une cassure double-brin que les autres allèles. Généralement, la réparation par conversion génique conduit au remplacement de l'allèle AT par une version enrichie en GC le cas échéant. De ce fait, le système NER pourrait lui aussi être indirectement impliqué dans le gBGC, bien que cela n'ait jamais été démontré explicitement [103]. Les régions génomiques possédant un fort taux de recombinaison sont généralement sujettes à un fort gBGC [43, 103].

Chez certains organismes, le génome est organisé en régions possédant des taux différentiels de GC. L'organisation des génomes de Vertrébrés en isochores (définis en régions riches en AT ou en GC) serait majoritairement issue des biais mutationnels qui influencent la composition nucléotidique de longues régions d'ADN [43, 106, 107]. La recombinaison génique serait l'un des facteurs les plus importants pour expliquer une telle mosaïque à l'échelle des génomes observés [107]. Au sein des isochores, le CUB des gènes est corrélé au contenu en GC des régions non-codantes (*i.e* régions flanquantes et introns) [43]. De ce fait, et à titre d'exemple, le CUB des

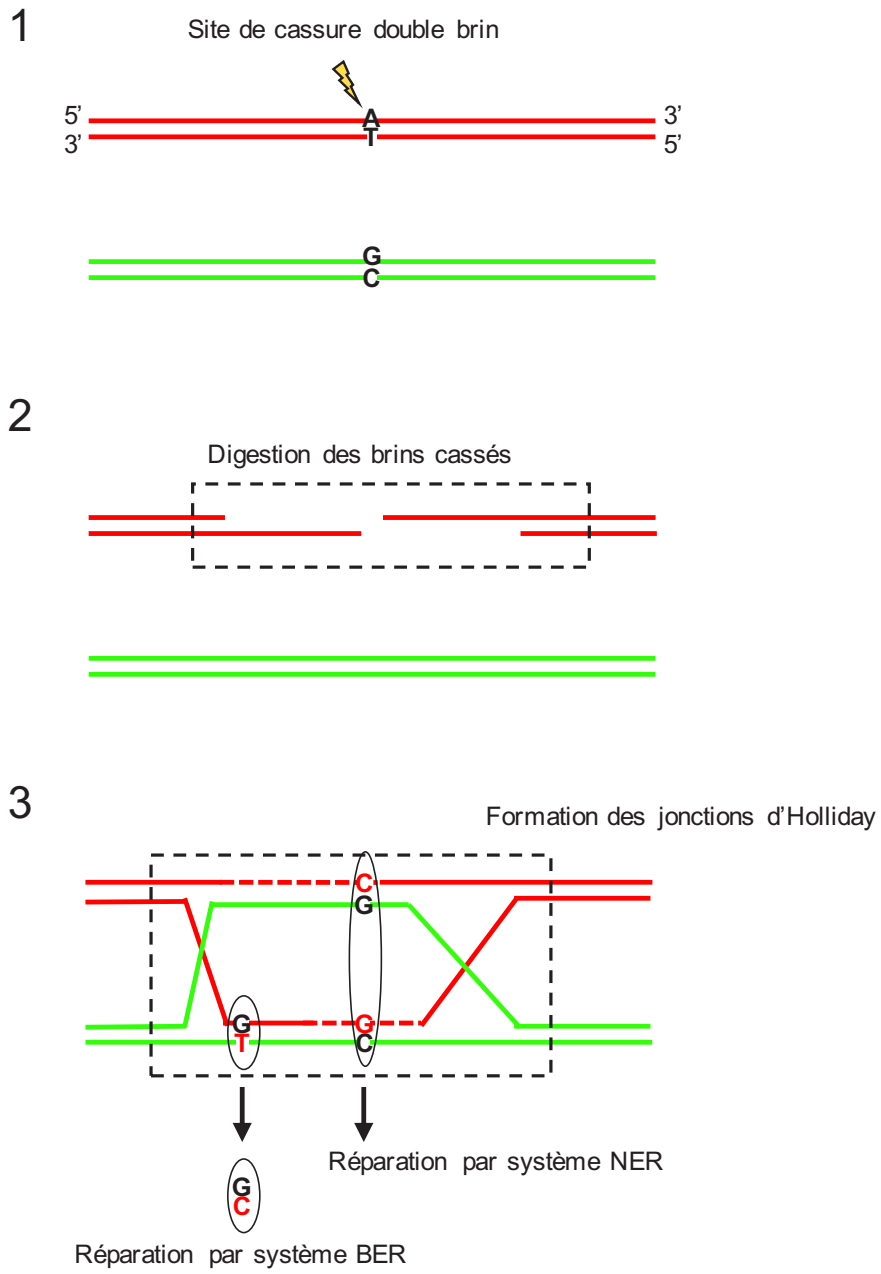

Figure 1.6 – **Description schématique du biais-GC de conversion génique (gBGC).** Étape 1 : l'une des deux chromatides subit une cassure double-brin ; Étape 2 : le site de cassure est rendu simple-brin sur les brins transcrits et non-transcrits par l'action de dégradosomes ; Étape 3 : les fragments monocaténaires de la chromatide brisée forment des hétéroduplexes avec la chromatide sœur. Le système NER répare les fragments monocaténaires en se servant de la chromatide sœur comme modèle. Le système BER répare les irrégularités entre les deux chromatides par substitution des nucléotides non-apparentés. Chacun des systèmes peut, selon la nature de la substitution (BER) et de la chromatide sœur (NER) modifier la composition nucléotidique de l'une ou des deux chromatides.

gènes humains est séparé en deux populations distinctes : une où le CUB est AT3-riche (au sein des isochores AT-riches) et l'autre où le CUB est GC3-riche (au sein des isochores GC3-riches) [94]. Quelle que soit l'origine de la composition nucléotidique observée, celle-ci ne s'arrête pas aux régions non-codantes, et définit fortement le CUB des gènes [43].

1.3 Cas particulier du CUB des virus

1.3.1 Définition du virus

Les virus sont les entités du vivant – bien que la définition du vivant reste toujours le sujet d'un vif débat – les plus abondantes. Ils possèdent un rôle majeur dans le façonnement de nos environnements et dans l'évolution de la biosphère [108]. On peut considérer que chaque organisme cellulaire est en étroite relation avec plusieurs virus qui, par une définition assez réductrice, constituent des parasites encapsidés intracellulaires de leur hôte [108]. On suppose que la relation entre les organismes cellulaires et les virus est ancienne, pour ne pas dire contemporaine de l'origine du vivant. Pour cette raison, mais aussi du fait de la vaste ubiquité de virus présents sur Terre, il est difficile d'estimer leur origine et leur histoire évolutive, en supposant même qu'ils soient monophylétiques [108]. Il existe trois scénarios distincts concernant l'origine des virus :

— L'hypothèse du virus primordial, où les virus seraient les descendants directs des premiers réplicons (*i.e.* un ARN ou un ADN capable de réplication) qui ont existé lors de l'étape pré-cellulaire des organismes que nous connaissons aujourd'hui [108].

— L'origine réductrice des virus, où les virus seraient en réalité le produit de la dégénérescence de cellules ayant perdu leur capacité d'auto-réplication, les forçant alors à devenir des parasites intra-cellulaires pour effectuer leur réplication [108].

— Le scénario des gènes échappés, où les virus sont apparus de multiples fois au cours de l'évolution par des mécanismes d'isolement de gènes « égoïstes » de la cellule, et qui ont gagné la capacité de se répliquer et d'infecter d'autres cellules de manière autonome. Au sein de cette hypothèse, on suppose que les virus d'eucaryotes, de bactéries et d'archées proviennent donc tous d'un génome primordial des hôtes qu'ils infectent [108].

Ces trois scénarios semblent distincts, exclusifs et voire même contradictoires, mais la diversité des virus que l'on observe aujourd'hui et les épisodes probables d'émergences non monophylétiques nous forcent à les considérer toutes les trois [108].

Selon la classification de Baltimore, les virus peuvent être classés par la nature de leur matériel génétique et de leur stratégie de reproduction : il existerait alors les virus à ARN (simples brins, formant les groupes IV, V et VI ; doubles brins, formant le groupe III) et les virus à ADN (simples brins, formant le groupe II ; doubles brins, formant les groupes I et VII) [109]. Le génome des virus est encapsulé au sein d'une capside (structure protéique protégeant le virus), elle-même parfois enveloppée par une membrane d'origine cellulaire et de protéines d'origine virales et cellulaires [110]. Grossièrement, les mécanismes de reproduction d'un virus orbitent autour du besoin de produire des virions pour infecter de nouveaux hôtes. Ainsi, la plupart des gènes viraux visent à promouvoir la réplication et la transcription du matériel génétique viral, et à produire les protéines de la capside.

Les virus sont des parasites intracellulaires obligatoires qui utilisent la machinerie cellulaire de leur hôte pour se reproduire [108, 110]. Le matériel génétique du virus ne fait pas exception dans le monde du vivant : il doit être répliqué, parfois transcrit et la production de ses protéines doit être assurée par une machinerie ribosomale. Les virus à ARN (à l'exception des rétrovirus du groupe VI) encodent leur propre ARN polymérase ARN-dépendante, qui permet la transcription de leur matériel génétique [111]. Les rétrovirus, quant à eux, codent une ADN polymérase ARN-dépendante, permettant la rétrotranscription de l'ARN viral en ADN. Ce dernier s'insèrera au sein du génome de la cellule et sera transcrit par les polymérases de l'hôte, permettant ainsi une réplication indirecte de l'ARN viral [111]. La grande majorité des virus à ADN encodent leur propre ADN polymérase (ou des enzymes qui contribuent à la réplication), et la transcription de leurs gènes est principalement assurée par des ARN polymérases de l'hôte, à l'exception des poxvirus qui, ne pouvant pénétrer le noyau de leur cellule hôte, utilisent une transcriptase d'origine virale. Les *Parvoviridae*, *Hepadnaviridae* et *Papovericetes* sont une exception chez les virus à ADN : les *Parvoviridae* et *Papovericetes* utilisent les ADN polymérases de leur hôte pour se répliquer, et les *Hepadnaviridae* se répliquent par le biais d'un ARN intermédiaire grâce à une transcriptase inverse virale [111]. Cette forte diversité dans les mécanismes de réplication et de transcription virales ne se retrouve pas au niveau du processus de traduction, où l'ensemble des ARN viraux sont traduits par la machinerie cellulaire de leur hôte [111].

1.3.2 Virus et CUB

Au sein des virus, on observe une véritable diversité dans leur CUB, où celui-ci peut être lié (ou non) à des processus de sélection traductionnelle [112–114]. Comme chez les organismes cellulaires, une sélection traductionnelle chez les virus impliquerait une meilleure traduction des gènes, et donc une amélioration de leur *fitness* (valeur sélective en français) [112, 115, 116].

Les bactériophages (*i.e.* virus de bactéries) semblent posséder un consensus dans leur CUB, où celui-ci suit généralement le CUB des gènes fortement exprimés de leur hôte [112, 117]. Mais il existerait tout de même une différence de CUB entre les gènes « précoces » (exprimés au début de l'infection de la cellule par le virus) et « tardifs » (exprimés à la fin de l'infection) chez un certain nombre de bactériophages [112, 118]. D'après *Goz et al.*, qui étudient le CUB des gènes du bactériophage λ d'*E. coli*, cette différence s'expliquerait par les changements de l'environnement cellulaire au cours des différentes phases du cycle lytique du virus (et notamment des populations d'ARNt), où les gènes précoces et tardifs seraient tour à tour « optimisés » pour exploiter au mieux la machinerie cellulaire de leur hôte [112].

Des conclusions identiques peuvent être tirées chez certains virus d'eucaryotes photosynthétiques, bien que la relation hôte-parasite du CUB dépend fortement du virus et de l'hôte considérés. Le CUB du virus otV5 de l'organisme *Ostreococcus tauri* semble être soumis à une pression de sélection lui permettant une « optimisation » de l'expression de ses gènes suivant l'hypothèse de sélection traductionnelle [116]. Il est intéressant de noter que ce virus possède cinq gènes codants des ARNts « complémentaires » de ceux retrouvés chez leur hôte. Ces ARNt sont peu représentés au sein de l'hôte, et la présence de ces gènes au sein du génome viral pourrait être liée à

un mécanisme assurant l'optimisation du CUB viral par l'apport d'ARNt supplémentaires [116]. L'analyse des phytovirus CTV (*Citrus Tristeza Virus*) sur trois espèces d'agrumiers a conduit à la détection d'une correspondance du CUB avec celui de l'une des trois espèces hôte [119]. Les infections sur l'espèce hôte partageant un fort CUB avec le CTV sont systématiquement asymptomatiques, alors que les deux autres présentent des symptômes allant jusqu'à l'apoptose des cellules et la mort de la plante. Cette correspondance du CUB pourrait marquer une avirulence progressive chez l'espèce hôte naturelle, et une virulence marquée chez les espèces devenues récemment hôtes du virus [119]. Pour *Biswas et al.*, une telle observation pourrait supposer un rôle du CUB dans la pathogénécité et dans la *fitness* des virus chez les plantes [119].

On observe chez un certain nombre de virus de Vertébrés à ARN un CUB discret (*i.e* proche d'une absence de biais), comme chez le virus du Zika [120], du Chikungunya [121] de Marburg [122] ou encore Ebola [123]. D'après une analyse extensive du CUB des virus à ARN par *Jenkins et Holmes*, il semblerait que le biais mutationnel soit majoritairement responsable du profil observé [114]. Mais les hypothèses de sélection traductionnelle ne sont pas pour autant rejetées. Au sein de leur étude sur le virus du Chikungunya ou encore de Marburg, *Butt et al.* ainsi que *Nasrullah et al.* démontrent clairement l'importance du biais mutationnel mais décrivent aussi la possibilité de l'existence d'une force de sélection actant sur les gènes des virus analysés [121, 122]. Il est intéressant de noter que certains virus à ARN, comme le virus de l'hépatite A ou encore de la grippe A possèdent un CUB contraire à celui de leur hôte et qui ne semble pas forcément et uniquement lié à des mécanismes mutationnels [124, 125]. Dans un autre exemple, on observe un CUB proche de celui de l'hôte *H. sapiens* chez le virus de la fièvre hémorragique de Crimée-Congo [126].

Le CUB des virus à ADN suit les mêmes tendances que celui des virus à ARN, où le profil des CDS est généralement marqué par un faible CUB, comme chez la majorité des représentants des *Parvoviridae* [113], chez le virus de l'hépatite B [127] ou encore chez les *Adenoviridae* [128]. Il est à noter que selon l'étude de *Shi et al.*, certains virus faisant partie des *Parvoviridae* ont un CUB fort et orienté vers celui de leur hôte, démontrant une diversité en CUB à l'intérieur d'une même famille [113]. Les *Papillomaviridae* ont eux aussi un fort CUB allant dans le sens contraire de celui de leur hôte [129]. Tout comme chez certains bactériophages, le virus de Epstein-barr possède une particularité remarquable où le CUB des gènes du cycle latent est contraire à celui de leur hôte, alors que le CUB des gènes associés à la phase lytique se démarquent par une proximité avec le CUB de l'hôte [130]. Il est à noter que *Karlin et al.* proposent qu'un CUB contraire à celui de l'hôte pourrait signifier une minimisation des conséquences délétères de l'infection. Bien entendu, ils appliquent cette hypothèse aux gènes latents du virus de Epstein-Barr, mais proposent aussi que le profil de ces gènes pourrait être dû à un événement d'acquisition récente de ces mêmes gènes [130]. Une étude récente de *Chen et al.* a démontré que le CUB des poxvirus, s'il suit celui de leur hôte, appauvrit les populations d'ARNt disponibles pour la cellule et entraînerait alors la dégradation de la production des protéines de l'hôte [131]. *Chen et al.* remarquent chez les virus de mammifères un rapprochement entre similarité du CUB virus-hôte et effet symptomatique de l'infection [131]. Pour eux, le CUB des virus de mammifères aurait tendance à ne pas suivre celui de leur hôte naturel et/ou réservoir, dans l'optique de pouvoir se transmettre sans provoquer de forts symptômes chez l'hôte. De ce fait, on pourrait considérer

que certains virus adopteraient une stratégie de « désoptimisation » pour alléger l'impact de l'infection sur l'hôte au cours de l'évolution [131].

Dans la majorité de ces études, le biais mutationnel semble être l'origine majoritaire du CUB observé chez les virus d'eucaryotes animaux, mais des signatures de sélection actant sur le CUB ont tout de même été observées, ou du moins supposées [113, 121, 122, 126, 127].

Comme énoncé précédemment, le CUB des virus est aussi sous l'influence des mécanismes de défense de la cellule hôte. En plus d'être soumis à la méthylation du matériel génétique viral, les dinucléotides CpG peuvent être reconnus par des systèmes de dégradation du matériel viral comme le ZAP ou encore les *Toll-Like Receptors* (Récepteur de Type Toll en français) [85, 132]. Certains virus, comme le BKPyV, évitent systématiquement l'usage de codons NCG pour coder les acides aminés correspondants, et ce probablement par une pression de sélection orchestrée par la cellule hôte (voir Chapitre 4). Une telle particularité a été explorée pour proposer de nouvelles solutions vaccinales [85, 115].

Il est primordial d'étudier les particularités génétiques des virus pour déterminer les origines de leur CUB. Au sein de cette thèse, nous étudierons le CUB des polyomavirus humains pour mieux comprendre la relation qu'ils entretiennent avec leur hôte, et pour tenter de déceler des signatures mutationnelles ou de sélection.

1.4 Mesure du biais d'usage des codons

1.4.1 Indices du CUB

Les indices de mesure du CUB visent principalement à représenter les fréquences des codons synonymes tout en réduisant l'information pour la rendre plus simple d'interprétation. Les indices de mesure du CUB peuvent être classés en deux familles qualifiant les indices de « comparatifs » et « non-comparatifs ». Le Tableau 1.3 décrit une liste non-exhaustive d'indices et les classe selon la famille à laquelle ils appartiennent.

Les indices comparatifs

Les indices comparatifs estiment le CUB en comparant une séquence requête à un jeu de données de référence [47]. Ces indices ont pour principal objectif de dénoter des différences ou des similitudes entre deux jeux de données, bien souvent pour faire ressortir une « optimalité » chez l'un d'entre eux, ou tout simplement pour vérifier si ceux-ci partagent le même CUB. Comme pour tout indice de mesure du CUB, le premier jeu de données correspond aux séquences requêtes. Selon l'indice et l'analyse voulue par l'utilisateur, le jeu de données de référence peut être le CUB d'un ensemble de gènes ou d'une liste de codons dits « optimaux » [37, 47]. Le premier type de référence est appelé Table d'Usage des Codons (TUC ; CUT en anglais), généralement au format de *Kazusa*, et contient des informations sur le nombre de codons retrouvés au sein de l'ensemble de gènes étudiés [133]. À l'instar du CUB, la TUC peut être définie à différents niveaux, comme sur la globalité des gènes (ou, par exemple, des gènes fortement exprimés) d'un organisme ou d'un tissu. Des indices tels que le CAI (pour *Codon Adaptation Index*), ou

TABLE 1.3 – **Liste non-exhaustive des mesures de biais d'usage des codons décrites au sein de cette thèse.** La nature de chaque méthode (indices comparatifs ou non-comparatifs), la nécessité d'une référence pour effectuer le calcul, l'intégration ou non du biais mutationnel dans l'analyse, les bornes des valeurs obtenues en sortie (allant d'un usage équiprobable des codons à un biais d'usage maximum, ou d'une « désoptimisation » à une « optimisation ») ainsi que la méthode d'origine sont autant d'informations indiquées dans les cases correspondantes.

Méthode	Auteurs	Famille	Bornes	Indice d'origine
FOP	*Ikemura* [37]	Comparative (réf : liste de codons optimaux)	$[0, 1]$	-
CAI	*Sharp et Li* [47]	Comparative (réf : jeu de séquences de référence)	$[0, 1]$	-
tAI	*Reis et al.* [135]	Comparative (réf : jeu de séquences de référence)	$[0, 1]$	CAI
$rCAI$	*Lee et al.* [136]	Comparative (réf : jeu de séquences de référence)	$[0, 1]$	CAI
χ^2	*Shields et al.* [137]	Non-comparative	$]-\infty, \infty[$	-
$CBI93$	*Morton* [138]	Non-comparative	$[0, 1]$	CBI [38]
$ICDI$	*Freire-Picos et al.* [134]	Non-comparative	$[0, 1]$	-
MCB	*Urrutia et Hurst* [139]	Non-comparative	$]-\infty, \infty[$	-
$SCUO$	*Wan et al.* [140]	Non-comparative	$[0, 1]$	-
CDC	*Zhang et al.* [141]	Non-comparative	$[0, 1]$	-
ENC	*Wright* [142]	Non-comparative	$[61, 20]$	-
ENC'	*Novembre* [143]	Non-comparative	$[61, 20]$	ENC
ENC^*	*Fuglsang* [144]	Non-comparative	$[61, 20]$	ENC
$mENC'$	*Satapathy et al.* [145]	Non-comparative	$[61, 20]$	ENC

encore l'ICDI (pour *Intrinsic CoDon bias Index*), utilisent une référence sous ce format [47, 134]. Le deuxième type de référence, simplement nommé liste des codons optimaux, représente une liste de codons synonymes qui sont considérés comme « optimaux » [37]. De ce fait, la liste représente pour chaque acide aminé un seul et unique codon synonyme « optimal ». L'optimalité d'un codon peut être définie en fonction de la disponibilité des ARNt associés, ou encore de la fréquence dudit codon par rapport à ses synonymes au sein de gènes fortement exprimés. Le FOP (pour *Frequency of Optimal codons*) [37] ou encore le CBI (pour *Codon Bias Index*) utilisent une liste de codons optimaux pour déterminer le CUB d'une séquence [38].

À ce jour, le CAI demeure l'indice comparatif le plus utilisé par la communauté scientifique. Celui-ci a pour objectif de comparer une séquence requête à une table d'usage des codons dans l'objectif de déterminer si les codons préférentiellement utilisés par la référence sont aussi utilisés au sein de la séquence requête et en quelle quantité [47]. De manière plus explicite, l'indice CAI mesure l'« optimalité » d'une séquence requête par rapport à la référence et n'est donc pas à proprement parler un indice de mesure du CUB (opinion personnelle).

La première étape de calcul d'un score CAI est la détermination de valeurs d'adaptativités relatives (w) pour chaque acide aminé (a) à partir des fréquences (f) des codons synonymes associés au sein de la référence. Ces valeurs représentent pour chaque codon c leur fréquence divisée par la fréquence de leur synonyme le plus représenté :

$$w_{c,a} = \frac{\mathrm{f}_a^c}{\mathrm{f}_a^{\max}} \tag{1.1}$$

De ce fait, les valeurs admises par les adaptativités relatives vont de 0 (absence du codon synonyme, si on ne considère pas la correction décrite ci-dessous) à 1 (score du codon synonyme le plus représenté). Dans le cas où un codon est absent de la référence, les auteurs de l'indice CAI suggèrent de lui donner une valeur d'adaptativité de 0,5 [47]. Le calcul de la valeur CAI se fait par la moyenne géométrique des adaptativités relatives des codons de la séquence requête selon la formule :

$$\mathrm{CAI} = \left(\prod_{k=1}^{L} w_k \right)^{\frac{1}{L}} \tag{1.2}$$

où k représente le $k^{\text{ième}}$ codon de la séquence requête et L sa longueur. Le score CAI d'une séquence requête varie entre 0 (utilisation par la requête des codons synonymes absents de la référence, sans correction des scores des adaptativités relatives) et 1 (utilisation exclusive des codons synonymes les plus représentés de la référence) [47].

Pour approfondir une analyse du CUB à l'aide de l'indice CAI, plusieurs versions de cet indice ont été créées au cours de ces dernières années. À titre d'exemple, le rCAI est une modification du CAI prenant en compte les trois cadres de lecture possibles d'une séquence contenant des gènes chevauchants [136]. Un tel indice a été développé dans le but d'analyser le CUB de séquences virales, où les régions codantes peuvent être exprimées selon différents cadres de

lecture (*i.e.*, les différentes façons de lire un CDS ; celui-ci peut être traduit par la machinerie ribosomale à partir du premier, deuxième ou troisième nucléotide du codon initiateur, décalant ainsi le reste des codons lus). *Lee et al.* proposent de soustraire à l'adaptativité relative w d'un codon c les adaptativités relatives que l'on retrouve en considérant les séquences de référence avec des cadres de lecture +1 et +2 [136] :

$$\text{LNWD}_c^a = \ln(w_{c,1}^a) - \frac{\ln(w_{c,2}^a) + \ln(w_{c,3}^a)}{2} \tag{1.3}$$

Pour une même séquence de référence, $w_{c,1}^a$ représente alors l'adaptativité relative d'un codon c obtenue selon le cadre de lecture classique, $w_{c,2}^a$ celle lorsqu'on décale le cadre de lecture d'un nucléotide et $w_{c,3}^a$ lorsqu'on le décale de deux nucléotides.

Une fois les valeurs LNWD obtenues, le calcul du score rCAI s'effectue, comme pour le CAI, en prenant la moyenne géométrique des valeurs LNWD des codons composant la séquence requête :

$$\text{rCAI} = \left(\prod_{k=1}^{L} \text{LNWD}_k \right)^{\frac{1}{L}} \tag{1.4}$$

Le tAI (pour *tRNA Adaptation Index*) est quand à lui une modification du CAI prenant en compte le nombre de copies d'ARNt au sein du génome de l'organisme étudié [135]. Ici, l'objectif est de représenter les hypothèses de sélection traductionelle et de disponibilité des ARNt lors de la traduction dans le calcul du CUB d'une séquence requête.

Tout d'abord, un score d'adaptativité absolue est calculé pour chaque codon selon la formule :

$$W_c^a = \sum_{j=1}^{n_c} (1 - s_j^c) * \text{tGCN}_j^c \tag{1.5}$$

où j représente l'ARNt associé au codon c, n_c le nombre total de copies de gènes des ARNt *cognate* et *near-cognate* du codon (nombre qualitatif d'ARNt pouvant s'associer à ce codon) du codon c, tGCN_j^c le nombre de copies de gènes codant l'ARNt j associé au codon c et s_j^c la contrainte sélective d'efficacité du couplage entre l'anticodon de j et le codon i.

Une fois toutes les adaptativités absolues obtenues, il est possible de calculer les adaptativités relatives de chaque codon :

$$w_c^a = \frac{W_c^a}{W_{\text{max}}} \tag{1.6}$$

où W_{max} représente l'adaptativité absolue la plus élevée de l'acide aminé a. Pour finir, le score tAI d'une séquence se calcule de la même manière que le score CAI, c'est à dire par la moyenne

géométrique des adaptativités relatives w des codons de la séquence requête :

$$\text{tAI} = \left(\prod_{k=1}^{L} w_k \right)^{\frac{1}{L}} \tag{1.7}$$

Au delà de l'indice CAI, l'indice FOP est lui aussi couramment utilisé pour déterminer le CUB d'une séquence requête [37]. Ici, l'objectif est de déterminer le nombre de codons « optimaux » (*e.g.* les codons associés à des ARNt possédant un grand nombre de copies de gènes) parmi tous les codons utilisés par la séquence requête :

$$\text{FOP} = \frac{Occ_{opt}}{N_{tot}} \tag{1.8}$$

Où Occ_{opt} représente l'occurrence des codons « optimaux » au sein de la séquence requête et N_{tot} le nombre total de codons.

Bien qu'ayant la même philosophie que le CAI, c'est-à-dire la mesure de l'« optimalité » d'un gène par rapport à une référence, le FOP n'intègre pas directement les codons synonymes « non-optimaux » dans son calcul, mais détermine plutôt la proportion d'optimalité d'une séquence.

Les indices non-comparatifs

La philosophie des indices non-comparatifs est de déterminer directement le CUB d'une séquence sans aucune référence et se rapproche donc de la description d'un véritable CUB. Ces indices sont réputées pour mesurer l'écart du CUB d'une séquence par rapport à un usage équiprobable des codons synonymes. Ici, Le CUB peut être déterminé à l'aide d'une Hypothèse Nulle (décrivant souvent ladite équiprobabilité des codons synonymes), d'une mesure de l'entropie des acides aminés en termes de CUB, ou encore en comparant les fréquences des codons de la requête face à des fréquences attendues [134, 142]. Cette famille contient des indices tels que le *Maximum likelihood codon bias* MCB [139], le *Synonymous Codon Usage Order* SCUO [140, 146] ou encore le χ*-scaled* [137].

L'ENC ou N_c (pour *Effective Number of Codons* ou le Nombre de codons Effectif en français) est l'indice le plus explicite et simple d'utilisation de cette famille. Il demeure aujourd'hui l'un des plus utilisés pour mesurer le CUB de séquences requêtes. En premier lieu, il est nécessaire de déterminer un score de contribution $\widehat{F_a}$ d'un acide aminé à partir des fréquences p_i de chaque codon synonyme pouvant le coder :

$$\widehat{F_a} = \left(n_a \sum_{i=1}^{k_a^{syn}} p_i^2 - 1 \right) (n_a - 1) \tag{1.9}$$

où n_a décrit l'usage total des codons pour l'acide aminé a (avec $n \in \{n_{1,a}, n_{2,a}, n_{3,a}, \ldots\}$), $p_i = n_i/n$ la fréquence du codon synonyme i et k_a^{syn} le nombre de codons synonymes de l'acide aminé a.

Si un codon est absent de l'analyse, il n'est pas pris en compte dans le calcul de la contribution $\widehat{F_a}$ de l'acide aminé auquel il est normalement attaché. $\widehat{F_a}$ possède une plage de valeurs allant de $1/k$ (les codons synonymes de l'acide aminé observé sont utilisés de manière égale au sein de la séquence) à 1 (fort biais d'usage des codons pour l'acide aminé étudié). Une fois les contributions $\widehat{F_a}$ calculées, il est possible pour chacune d'entre elle de calculer leur valeur $\widehat{\mathrm{N}}_\mathrm{c}$:

$$\widehat{\mathrm{N}}_{\mathrm{c},a} = \frac{1}{\widehat{F_a}} \tag{1.10}$$

La valeur du $\widehat{Nc}$ est directement corrélée au biais d'usage des codons de l'acide aminé étudié. Ainsi, $\widehat{Nc} = k$ si les codons sont utilisés de manière équivalente par l'acide aminé. $\widehat{Nc} = 1$ si, par exemple, un codon est exclusivement utilisé pour coder l'acide aminé.

Le calcul global de l'ENC de la séquence requête est effectué par l'addition des moyennes des contributions $\widehat{F}$ de chaque catégorie d'acide aminé (*i.e.* acides aminés encodés par 2, 3, 4 ou 6 codons). En considérant, par exemple, les acides aminés de type 4 (*i.e.* tous les acides aminés encodés par 4 codons synonymes) on obtient l'équation suivante :

$$\frac{x_4}{\frac{1}{x_4}\sum_{i=1}^{k}\widehat{F_i}} = \frac{5}{\bar{\bar{\widehat{F}}}_4} \tag{1.11}$$

où x_4 représente le nombre d'acides aminés encodés par 4 codons synonymes, $\widehat{F_i}$ la valeur des contributions de chacun de ces acides aminés, $\bar{\bar{\widehat{F}}}_4$ la moyenne des contributions des 5 acides aminés encodés par 4 codons synonymes et k le nombre de codons synonymes codant l'acide aminé a.

Pour finir, on peut calculer l'ENC global de la séquence d'intérêt :

$$\widehat{\mathrm{N}}_\mathrm{c} = 2 + \frac{9}{\bar{\bar{\widehat{F}}}_2} + \frac{1}{\bar{\bar{\widehat{F}}}_3} + \frac{5}{\bar{\bar{\widehat{F}}}_4} + \frac{3}{\bar{\bar{\widehat{F}}}_6} \tag{1.12a}$$

$$= \sum_{i \in ARC} \left(\frac{x_i}{\bar{\bar{\widehat{F}}}_i}\right) \tag{1.12b}$$

le score obtenu varie de 61 (tous les codons synonymes sont utilisés avec la même fréquence) à 20 (tous les acides aminés de la requête sont codés par un seul et unique codon synonyme).

À l'instar du CAI, l'ENC a aussi été sujet à plusieurs modifications visant cette fois-ci à améliorer la fiabilité de son calcul. En effet, le score de l'ENC peut être modifié et bien souvent surestimé en fonction de l'absence de certains acides aminés dans la séquence étudiée. Pour corriger ce biais, *Satapathy et al.* ont développé un indice appelée $m\widehat{\mathrm{N}}_\mathrm{c}$. En premier lieu, il est nécessaire de calculer les contributions de chaque acide aminé de l'analyse :

$$\widehat{F}'_a = \frac{\chi_a^2 + 1}{k^a_{\mathrm{syn}}} \tag{1.13}$$

où χ^2_a est une statistique de χ^2 sur l'acide aminé a :

$$\chi^2_a = \sum_{c=1}^{k^a_{syn}} \frac{(p^a_c - e^a_c)^2}{e^a_c} \tag{1.14}$$

p_c représente la fréquence observée du codon c de l'acide aminé a et e_c sa valeur attendue sous une hypothèse nulle prenant en compte un biais mutationnel défini [143, 145]. Une fois toutes les valeurs de contribution obtenues, le calcul du $\mathrm{m\widehat{N}_c^{'}}$ est proche de celui du $\widehat{\mathrm{N}}_\mathrm{c}$:

$$m\widehat{\mathrm{N}}_\mathrm{c}^{'} = \sum_{a \in \mathcal{A}} \frac{1}{F'_a} \tag{1.15}$$

où $\mathcal{A}$ représente tous les acides aminés présents dans la séquence requête.

Le calcul du $m\widehat{\mathrm{N}}_\mathrm{c}^{'}$ est donc globalement identique à celui du $\widehat{\mathrm{N}}_\mathrm{c}$, à l'exception du fait que le nombre de codons codant chaque acide aminé n'est plus pris en compte dans chacune des étapes de calcul [145]. Si la valeur attendue est égale à une hypothèse d'usage équivalent des codons, on retrouve une méthode « classique » proche de celle de Wright [142] avec :

$$\widehat{F}'_a = \sum_{c=1}^{k^a_{syn}} (p^a_c)^2 \tag{1.16}$$

et

$$m\widehat{\mathrm{N}}_\mathrm{c}^{'} = \sum_{a \in \mathcal{A}} \frac{1}{F'_a} \tag{1.17}$$

Comme pour le $\widehat{\mathrm{N}}_\mathrm{c}$, le score obtenu varie de 61 (tous les codons synonymes sont utilisés avec la même fréquence) à 20 (tous les acides aminés de la requête sont codés par un seul et unique codon synonyme).

Limites des indices de mesure du CUB

Malgré une grande diversité dans les stratégies de calcul du CUB, bon nombre d'indices sont sujets à des défauts et à des limites pouvant impacter la significativité du score obtenu. La première d'entre elle est la taille de la séquence qui, sous couvert de représenter tous les acides aminés, doit avoir une taille supérieure à 300 nucléotides pour que son CUB soit estimé convenablement [147]. En effet, en dessous de cette taille, il manque normalement à la plupart des séquences un ou plusieurs acides aminés, et les différents acides aminés qui les composent ne sont généralement pas assez représentés, ce qui peut fausser toute mesure de CUB [148]. Avec l'ENC, l'absence d'une famille de codons synonymes, et plus précisément celle de l'isoleucine (le seul acide aminé codé par trois codons synonymes), provoque un véritable changement des

plages de scores obtenues et, sans modification empirique du calcul, une surestimation du CUB [145]. C'est aussi le cas de l'indice CAI pour lequel le score des codons synonymes manquants à une référence a obligatoirement une valeur adaptative de 0.5, changeant alors le score des séquences requêtes sans que le véritable impact des codons synonymes dans le CUB ne soit estimé [47]. Dans ce cas, il serait plus judicieux de construire une meilleure référence, ou encore d'ignorer ces codons lors du calcul de la valeur CAI de la requête.

La composition et l'hétérogénéité en acides aminés d'une séquence peut aussi être un facteur limitant lorsqu'il s'agit de déterminer avec précision le CUB d'une séquence. Si l'on prend le cas d'une séquence requête virtuelle possédant un grand nombre d'asparagine et un petit nombre de phénylananine, le CUB de la phénylalanine pourrait être masqué par le CUB de l'asparagine. Comptabiliser les fréquences des acides aminés, c'est représenter le véritable CUB d'une séquence, mais c'est aussi masquer le CUB individuel de chaque acide aminé. Alors que l'indice CAI comptabilise tous les codons synonymes d'une séquence, et donc indirectement sa composition en acides aminés [47], l'indice ENC standardise le CUB de chaque acide aminé de manière à ce qu'ils aient le même poids dans le score final [142].

Pour finir, le contenu en GC des séquences requêtes peut lui aussi influencer le calcul du CUB d'une séquence. Toujours en prenant l'exemple de l'indice CAI, les scores globaux d'un organisme ayant un fort taux en GC ou en AT, et donc un fort biais d'usage des codons global, seront constamment élevés, biaisant alors les résultats obtenus et leur interprétation, surtout lorsque l'on compare ces scores à ceux d'un organisme ayant un taux de GC plus hétérogène [94].

Une autre problématique s'impose lorsqu'il s'agit de parler de détermination du CUB. Certains indices, comme l'ENC, déterminent directement le CUB d'une séquence de manière remarquable, et la signification des résultats possède une forte représentativité biologique. Mais d'autres indices, et notamment les comparatifs, sont souvent considérés comme déterminant du CUB alors qu'ils mesurent bien souvent une « optimalité » des codons par rapport à ladite référence. C'est le cas du CAI qui, malgré sa sémantique, est souvent utilisé pour mesurer le CUB d'une séquence requête en rapport avec une référence. Or, un score CAI élevé n'indique en rien le CUB d'une séquence, mais plutôt si les codons fortement utilisés dans la référence se retrouvent aussi en grand nombre dans la séquence requête [47].

Autres méthodes de mesure du CUB

Les indices décrits ci-dessus ont pour principal objectif de résumer l'information pour permettre une analyse aisée et synthétique du CUB. *De facto*, cette réduction conduit automatiquement à une perte de détails sur la nature du CUB. À titre d'exemple, deux séquences requêtes ayant une différence marquée dans leur CUB peuvent avoir un même score CAI, tout simplement parce que la combinaison des valeurs adaptatives de leurs codons s'équilibre autour d'une même valeur. Pour compléter de telles analyses, il est nécessaire de considérer les codons à l'échelle des différentes familles de codons synonymes auxquelles ils appartiennent, de manière à destructurer et à décrire le CUB d'une séquence acide aminé par acide aminé. En parallèle du développement de l'indice CAI, *Sharp et Li* développèrent le calcul des scores RSCU (*Relative Synonymous Codon Usage* ou valeurs relatives d'usage des codons en français) [47]. Avec le RSCU, les fréquences des codons synonymes sont transformées en fonction de leur relation avec

les fréquences des autres codons synonymes qui encodent le même acide. Le RSCU, bien qu'il ne possède aucun avantage par rapport à une simple représentation des fréquences des codons synonymes, est bien souvent préféré à cette solution car il permet une meilleure représentation du CUB. Celui-ci est calculé de la manière suivante :

$$RSCU_{c,a} = \frac{k_a^{syn} \times x_{a,c}}{\sum_{c=1}^{k_a^{syn}} x_{a,c}} \tag{1.18}$$

où k_a^{syn} représente le nombre de codons synonymes encodant l'acide aminé a et $x_{a,c}$ l'occurrence des codons synonymes de a. Ainsi, un codon synonyme associé à un acide aminé encodé par quatre codons synonymes peut avoir une valeur RSCU allant de 0 (aucune utilisation du codon) à 4 (utilisation unique du codon).

Des modifications théoriques au calcul du RSCU ont été proposées par certains auteurs, tel que le RSCUrs de *Paulet et al.* [149]. Celui-ci se base sur la détermination du profil ribosomal des ARNm fortement exprimés d'un organisme pour en déduire le RSCU des codons. De ce fait, cette méthode s'appuie sur la sélection traductionnelle pour représenter un CUB en fonction du succès traducteur qu'il apporte. Les auteurs soulignent par ailleurs que le RSCUrs prédit et représente correctement l'expression des gènes chez une pléthore d'organismes, mais que son importance est plus limitée chez les eucaryotes multicellulaires [149]. Le calcul du RSCUrs est le même que pour le RSCU, à l'exception du fait que les codons comptabilisés dans le calcul du RSCU appartiennent à des gènes fortement exprimés et sont situés de la position 20 à 200 du gène et ce pour éviter tout autre biais d'accumulation des ribosomes, comme indiqué précédemment avec la mise en place d'une « rampe » de lancement des ribosomes au sein de la région 5' des gènes [149].

1.4.2 Outils de mesure du CUB

En parallèle du développement de nouvelles mesures du CUB, plusieurs outils informatiques ont été créés. Ceux-ci permettent aux utilisateurs de calculer directement le CUB de leurs séquences requêtes à l'aide d'un certain nombre d'indices tout en proposant diverses tâches pour approfondir leurs analyses. Ces programmes possèdent une architecture simple où l'utilisateur saisit en entrée les séquences requêtes (ainsi qu'une référence si besoin) et choisit des options dépendantes de la tâche demandée, pour recevoir en sortie les résultats des analyses (Figure 1.7). Une liste non-exhaustive des programmes et outils informatiques de mesure du CUB est présentée au sein du Tableau 1.4.2.

Bien souvent, ces programmes se contentent d'effectuer les mesures à l'aide des indices CAI, ENC ou d'autres indices mais sans jamais proposer une liste exhaustive de solutions. À ce jour, l'outil `CodonW` reste le plus complet en proposant une détermination du CUB par le biais des indices CAI, ENC, CBI et ICDI [150]. Il est à noter que certains outils ont été spécialement développés pour y insérer un index, comme l'outil CodonO pour l'index SCUO [146]. *A contrario*, certains indices, tels que le MCB et le χ^2 n'ont jamais été implémentés au sein d'un outil disponible pour la communauté scientifique, et peuvent donc être considérés comme « orphelins ». Un tel manque dans la disponibilité des solutions de calcul du CUB est conséquent, ne serait-ce que

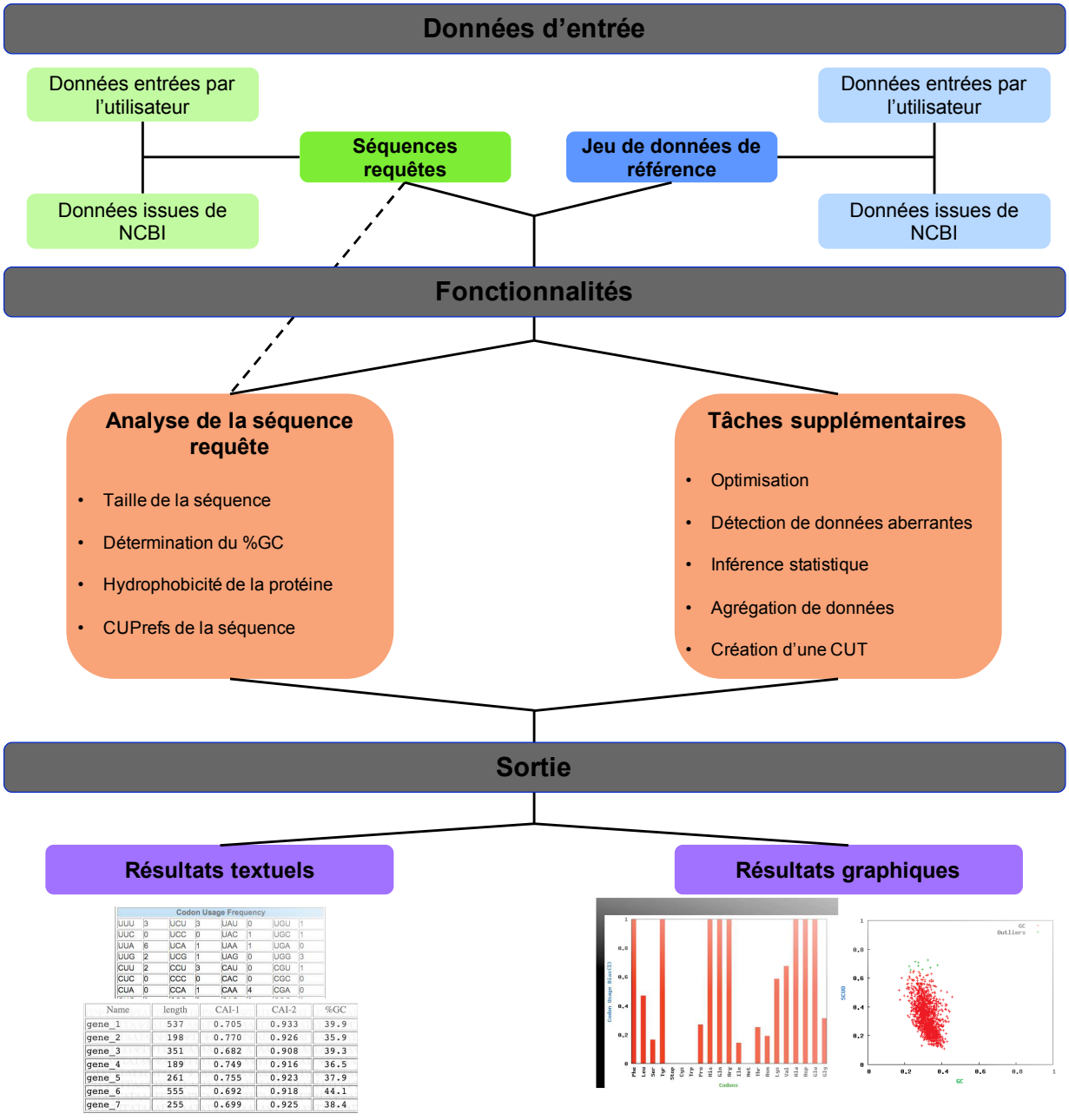

Figure 1.7 – **Architecture type d'un outil d'analyse du biais d'usage des codons.** Pour fonctionner, ces programmes demandent un certain nombre de données d'entrée, telles que des séquences requêtes (obligatoire), une table d'usage des codons au format de *Kazusa* (facultatif selon l'outil) ou d'autres entrées spécifiques aux fonctionnalités proposées [151, 152]. Plusieurs tâches sont ensuite effectuées telles que la détermination de la taille, du contenu en GC ou encore du CUB des séquences requêtes par le biais des différents indices contenus. Des tâches facultatives mais permettant un approfondissement du CUB peuvent être effectuées selon l'outil, telles qu'une optimisation des séquences, une détection de données aberrantes (« *outliers* ») ou encore la création d'une table d'usage des codons [146, 151]

TABLE 1.4 – **Liste non exhaustive des programmes et outils informatiques d'analyse du biais d'usage des codons.** Le type d'outil (léger ou « standalone »), les données prises en entrée, la nature des données en sortie, les mesures implémentées et pour finir les fonctionnalités spécifiques de chaque outil sont présentée au sein de ce tableau.

Outil	auteurs	Type d'outil	Données en entrée	Données en sortie	Mesures implémentées	Fonctionnalités spécifiques
`CAIcal`	*Puigbò et al.* [152]	Outil en ligne	séquences requêtes (fasta) / table d'usage des codons (kazusa)	textuelles et graphiques	CAI, ENC	e-CAI / Optimisation
`CodonO`	*Angellotti et al.* [146]	Standalone	séquences requêtes (fasta) / table d'usage des codons (kazusa)	textuelles et graphiques	SCUO	Détection de gènes outliers / nuages de points
`CodonW`	*Peden et Sharp* [150]	Standalone	séquences requêtes (fasta) / table de fréquence des codons (« .coa »)	textuelles	CAI, CBI, FOP, ENC	-
`JCAT`	*Grote et al.* [151]	Outil en ligne	séquences requêtes (fasta) / table d'usage des codons (kazusa)	textuelles et graphiques	CAI	Optimisation
`ACUA`	*Vetrivel et al.* [153]	Standalone	séquences requêtes (fasta) / table d'usage des codons (kazusa / « .cut »)	textuelles	CAI, ENC	-
`EMBOSS`	*Rice et al.* [154]	Suite de programmes	séquences requêtes (fasta) / table d'usage des codons (kazusa / « .cut »)	textuelles	CAI	-
`INCA`	*Supek et Vlahovicek* [155]	Standalone	séquences requêtes (fasta / « .ffn ») / table d'usage des codons (kazusa)	textuelles	CAI, ENC	Clustering / optimisation / nuages de points

pour les analyses de *benchmark* (banc d'essai en français) qui comparent l'efficacité de différents indices à représenter le CUB.

Fonctionnalités des programmes de mesure du CUB

Certains outils informatiques et logiciels proposent des tâches supplémentaires pour améliorer l'analyse du CUB, par exemple par le biais d'une organisation des séquences requêtes selon leur CUB, ou par leur transformation pour qu'elles suivent un CUB particulier.

Vu l'importance du CUB dans l'expression des gènes bactériens, une fonctionnalité d'optimisation est bien souvent mise en avant par ces outils. Celle-ci a pour objectif de modifier le CUB d'une séquence pour qu'il suive celui d'une référence particulière (Figure 1.8). Majoritairement utilisée pour optimiser les gènes bactériens, cette fonctionnalité est prisée en bioingénierie pour augmenter l'expression de gènes cibles [156]. Plusieurs optimisations sont alors possibles :

— *One amino acid - one codon*, où seul un des codons synonymes d'un même acide aminé est représenté au sein de la séquence.

— *Random-guided*, où chaque codon synonyme d'un acide aminé a une probabilité d'être sélectionné en fonction de sa fréquence au sein d'un jeu de données de référence.

— *Random*, où chaque codon synonyme a la même probabilité d'être sélectionné.

Pour chacune de ces approches, l'outil `OPTIMIZER` propose d'orienter l'optimisation vers un CUB GC-enrichi ou AT-enrichi [157]. Pour aller plus loin dans l'optimisation des séquences, le programme `JCAT` permet d'éviter la création de sites de restriction lors de l'optimisation des séquences [151]). Il est en effet important de prendre en compte la stabilité des ARNm lors de leur optimisation, notamment au travers de la formation de structures secondaires.

Puigbò et al. ont proposé une approche statistique de la mesure du CUB sur indices comparatifs avec l'outil `e-cai` [152, 158]. Ici, l'objectif est de simuler un certain nombre de séquences avec la même composition en acides aminés que la séquence requête, mais en optimisant le CUB selon une approche *random-guided*. De ce fait, les séquences obtenues suivent la référence avec une certaine liberté relative aux fréquences des codons synonymes. Une simple comparaison d'un score CAI de la séquence requête à la distribution des scores des séquences simulées permet ainsi une inférence statistique du CUB [152, 158]. En utilisant ce principe, ainsi qu'un jeu de données de référence pré-établi consistant en un ensemble de gènes fortement exprimés, `e-cai` peut prédire si une séquence requête appartenant au génome d'*E. coli* est fortement exprimée au sein de cet organisme [152, 158].

Le programme `INCA` propose une fonctionnalité d'agrégation des séquences requêtes selon leur CUB [155]. Pour ce faire, cet outil utilise un algorithme non supervisé de type *neural-network* (de réseau neuronal en français) appelé *Self Organizing Map* (MOP ; Carte auto-organisée en français) qui permet d'effectuer une analyse de cette nature sur, par exemple, les fréquences des codons synonymes des séquences requêtes. Une telle fonctionnalité demeure primordiale lorsqu'il s'agit d'analyser un grand nombre de séquences dans l'objectif de déterminer si leur CUB se démarque de celui des autres.

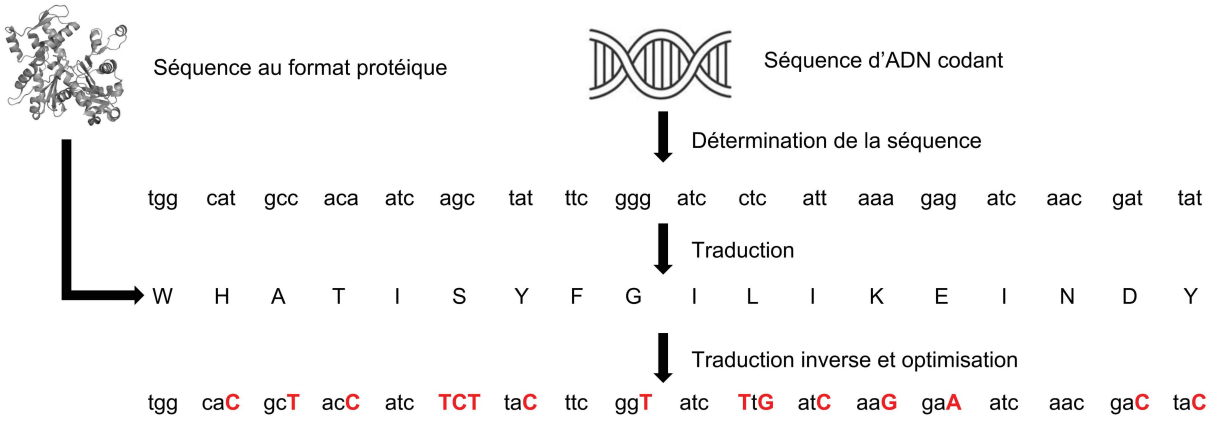

Figure 1.8 – **Architecture type d'une fonction d'optimisation des gènes.** L'utilisateur doit tout d'abord fournir une séquence d'ADN codante (qui subit alors une étape de traduction en séquence protéique) ou une séquence protéique. Pour toute optimisation, il est nécessaire d'insérer un jeu de données de référence (non présenté dans ce schéma). Pour chaque acide aminé de la séquence protéique obtenue, l'étape d'optimisation choisit un acide aminé selon une probabilité dictée par la table de référence : il s'agit donc d'une étape de traduction inverse avec élimination des codons rares. La séquence nucléotidique obtenue est renvoyée à l'utilisateur. Sur ce schéma, les nucléotides modifiés sont en majuscule. Cette figure est issue de l'article de *Richardson et al.* [156].

1.5 Les polyomavirus humains

1.5.1 Généralités

La découverte du premier polyomavirus remonte au milieu des années 1950, lors de l'analyse d'une souris présentant de multiples tumeurs au niveau des glandes salivaires [159, 160]. L'origine de cette affliction a d'abord été supposée d'ordre chimique, mais après une analyse poussée, il a été déduit que l'apparition des tumeurs était due à une infection par un nouveau virus, le polyomavirus murin (MPyV) [159, 160]. L'étymologie du nom est plutôt explicite : elle provient du grec « πολλοι » (plusieurs) et « ωμα » (tumeur) [161]. Avant la découverte d'un grand nombre de polyomavirus au tropisme cellulaire et aux stratégies d'infection divergents, il était en effet supposé que ceux-ci provoquaient systématiquement l'apparition de multiples tumeurs chez leur hôte [159, 160].

Un autre grand pas dans la découverte et la compréhension des polyomavirus a été fait au cours d'une histoire sordide : le premier polyomavirus simien, SV40 (pour *Simian Virus 40* ou Virus simien 40 en français, aussi appelé virus vacuolant simien 40) a été isolé en 1960 lors de l'analyse de vaccins contre la polio issus d'une production à partir du macaque rhésus *Macaca mulatta* [162]. Pire, ce virus n'aurait pas été rendu inactif au sein des vaccins administrés car plus résistant que le poliovirus au traitement par le formaldéhyde [163]. Cela dit, la gravité de

cette contamination a été dans un premier temps pondérée par l'annonce que, bien que des effets cytopathiques de SV40 ont été observés chez le grivet *Chlorocebus aethiops*, aucun effet délétère n'a été remarqué chez le macaque rhésus. Mais ce scandale est revenu sur la table lorsque le caractère oncogène de SV40 a été démontré chez le hamster [164]. Dès lors, les effets délétères, voire oncogènes, du SV40 chez *H. sapiens* continuent à faire l'objet d'un débat, relevant la réputation de cet incident au niveau de celui du *Cutter incident* [165, 166]. Tout d'abord considéré comme un virus vacuolant (*i.e.* provoquant l'apparition de vacuoles dans la cellule infectée), le SV40 a été ensuite reconnu comme faisant partie des polyomavirus [162].

C'est en 1971 que les premiers polyomavirus humains ont été détectés. Les polyomavirus BK (BKPyV) et JC (JCPyV), nommés d'après les initiales des patients immunosupprimés chez qui ils ont été isolés pour la première fois, ont étés découverts à quelques mois d'intervalles [167, 168]. Puis, au fur et à mesure, de nouveaux polyomavirus humains ont été isolés, tels que le polyomavirus à cellules de Merkel (MCPyV) en 2008 [169], ou encore le polyomavirus de la trichodysplasie spinulosique (TSPyV) en 2010 [170]. À ce jour, 15 polyomavirus humains, dont l'ensemble infecte un large nombre de tissus, ont été découverts (Tableau 1.5). Certains d'entre eux peuvent notamment provoquer des infections symptomatiques plus ou moins grave chez les personnes immunosupprimées. C'est le cas des virus BKPyV, JCPyV, MCPyV et TSPyV : ceux-ci peuvent provoquer des cystites hémorragiques, des rejets de greffons de rein par néphropathie (PVAN ; BKPyV), des leucoencéphalopathies multifocales progressives (LEMP ; JCPyV), des cancers de la peau (MCPyV) ou des altérations bénignes de la peau (TSPyV) [169–172].

À ce jour, ce sont environ une centaine de polyomavirus qui ont été isolés chez une large gamme d'hôtes bilatériens tels que les arachnéens, les insectes, les poissons, les oiseaux, et, à une plus grande mesure, les mammifères [173]. Ces polyomavirus sont généralement considérés comme spécialistes d'un seul hôte et sont relativement peu virulents, sauf chez les oiseaux où, à titre d'exemple, le polyomavirus aviaire (APyV) infecte un grand nombre d'espèces différentes et provoque des inflammations bien souvent mortelles pour l'hôte [163]. Le niveau de spécialisation n'est pas exclusif chez les mammifères : certains polyomavirus peuvent en de rares cas infecter un hôte proche de l'hôte naturel, mais avec une affinité bien moindre [163].

Les polyomavirus possèdent de fortes similitudes avec les papillomavirus et ont formé avec ceux-ci la famille des *Papovaviridae* (pour **PA**pillomavirus - **PO**lyomavirus - **VA**cuolating virus) jusqu'à la fin des années 1990 (voir septième rapport de l'ICTV, ou *International Committee on Taxonomy of Viruses* [174]). En effet, il s'agit de virus possédant un ADN bicaténaire et circulaire, court et semblable en certains points. Ces virus sont non-encapsulés, possèdent une capside icosaédrale organisée en 72 pentamères et certains d'entre eux peuvent provoquer des lésions cutanés [172]. Cela dit, l'analogie s'arrête vite : les virions des papillomavirus sont plus imposants que ceux des polyomavirus (55nm contre 40-45nm) et l'organisation, la taille et le contenu de leurs génomes diffèrent tout autant que leur tropisme cellulaire. La révision de la classification des polyomavirus et des papillomavirus les font aujourd'hui appartenir à des familles distinctes (les familles *Polyomaviridae* et *Papillomaviridae*), mais ces virus sont toujours rassemblés sous la classe des *Papovaviricetes* [174].

Au sein de ce projet, nous allons focaliser notre recherche sur les polyomavirus humains, et plus précisément sur le BKPyV, connu pour provoquer le dysfonctionnement d'un greffon de rein par établissement d'une PVAN [171]. La PVAN apparaît suite à la mise en place de l'im-

TABLE 1.5 – **Tableau récapitulatif des différents polyomavirus humains connus à ce jour.** Ce tableau indique le nom du polyomavirus (et la référence de sa découverte), son abbréviation, la pathogénicité qu'il peut induire ainsi que le compartiment biologique dans lequel il est communément retrouvé. Cette table s'inspire de celle décrite au sein du doctorat de *Mazalrey* [172].

Souche	**Abb.**	**Pathogénicité**	**Compartiment biologique**
BKPyV [168]	BKPyV / HPyV1	PVAN, cystite hémorragique	système uro-génital
Polyomavirus JC [167]	JCPyV / HPyV2	LEMP	liquide céphalo-rachidien
Polyomavirus KI [175]	KIPyV / HPyV3	-	Système respiratoire
Polyomavirus WU [176]	WUPyV / HPyV4	-	Système respiratoire
Polyomavirus de Merkel [169]	MCPyV / HPyV5	Carcinomes à cellules de Merkel	Peau
Polyomavirus humain 6 [177]	HPyV6	-	Peau
Polyomavirus humain 7 [177]	HPyV7	-	Peau
Polyomavirus de la tricho. spinu. [170]	TSPyV / HPyV8	trichodysplasie spinulosique	Peau
Polyomavirus humain 9 [178]	HPyV9	-	Système uro-génital
Polyomavirus de Malawi [179]	MWPyV / HPyV10	-	Système digestif
Polyomavirus de Saint Louis [180]	STLPyV / HPyV11	-	Système digestif
Polyomavirus humain 12 [181]	HPyV12	-	Système digestif
Polyomavirus du New Jersey [182]	NJPyV/ HPyV13	Myosites	Cellules endothéliales du muscle
Polyomavirus Lyon-Iarc [183]	LIPyV / HPyV14	-	Peau
Polyomavirus du Québec [184]	QPyV / HPyV15	-	Système digestif ?

munosuppression chez des patients receveurs de la greffe pour empêcher un rejet par le système immunitaire. Sans système immunitaire compétent, le virus prolifère au niveau des cellules épithéliales des glomérules du rein et provoque une inflammation provoquant le dysfonctionnement du greffon. À ce jour, il n'existe pas de traitement efficace contre la prolifération des BKPyV. La seule méthode qui a fait ses preuves est la modulation de l'immunosuppression pour empêcher une prolifération virale incontrôlable, mais cette solution peut aussi provoquer le rejet du greffon. Pour mieux comprendre les mécanismes sous-jacents de la PVAN, il est nécessaire de mieux comprendre l'évolution intra-hôte des BKPyV, leur diversité et spécificités génomiques dans le cadre d'une infection sans système immunitaire (et donc avec un relâchement des pressions de sélection). Ces données pourraient nous aider à mieux comprendre pourquoi, dans certains cas et pas dans d'autres, les BKPyV provoquent la PVAN.

1.5.2 Génome des polyomavirus

Le génome des polyomavirus est sous la forme d'une molécule d'ADN circulaire et courte (environ 5000 à 5500 paires de base), double brin et circulaire. Au sein des génomes de polyomavirus, on retrouve une certaine diversité dans le contenu en gènes, bien que leur structure globale reste la même. Le génome peut être grossièrement séparé en trois régions. Une première, appelée NCCR (pour *Non-Coding Control Region* ou région de contrôle non codante en français), contrôle l'expression de deux ARNm qui, selon l'épissage qu'ils subissent, produisent les différents gènes viraux. Une deuxième comprend le premier ARNm, situé à la limite 5' de la région NCCR encode les gènes dits précoces. La troisième région contient elle aussi un ARNm qui est quant à lui responsable de la production des gènes dits tardifs [171] (Figure 1.9).

Il existe deux gènes précoces que l'on retrouve chez tous les polyomavirus. Ceux-ci s'expriment à partir du même ARNm dès l'entrée du virus dans la cellule, avant même que le génome viral ne commence à être répliqué [185]. On retrouve chez la majeure partie des polyomavirus les gènes *LT* (*Large tumoral Antigen* ou Grand antigène Tumoral en français), une version tronquée de *LT* ainsi que le gène *ST* (*Small Tumoral antigen* ou petit antigène tumoral en français). Ces gènes ont reçu l'étiquette de « tumoral » car ils étaient supposés être des initiateurs de la transformation des cellules chez les premiers polyomavirus découverts. Le gène *LT* se retrouve à proximité de la limite 3' de la région NCCR du génome. Il est divisé en deux exons séparés par un intron d'environ 350 paires de bases. Les fonctions associées au gène *LT* sont encore peu connues, mais plusieurs hypothèses ont été formulées au cours des années passées. Tout d'abord, le gène *LT* serait impliqué dans les mécanismes de la réplication virale : il bloque la cellule hôte en phase S (favorable à la réplication du matériel génétique), joue le rôle d'hélicase pour l'ADN viral et recrute des facteurs cellulaires pour promouvoir la réplication [186]. Enfin, et comme l'origine de son nom l'indique, la forme tronquée du gène *LT* de MCPyV peut induire une transformation de la cellule hôte en cellule cancéreuse par une affinité inhibitrice avec les protéines p53 et Rb [187]. La région précoce de MCPyV possède aussi un autre gène appelé *ALTO* (pour *Alternative Large T Open reading frame* ; cadre de lecture alternatif de l'antigène T en français), exprimé à partir d'un cadre de lecture différent de celui du gène *LT*, mais ayant putativement un rôle dans la réplication virale similaire à ce dernier gène [188]. Le gène *ST* partage une région commune avec le *LT* et sa fonction ne semble pas nécessaire pour le cycle de reproduction du

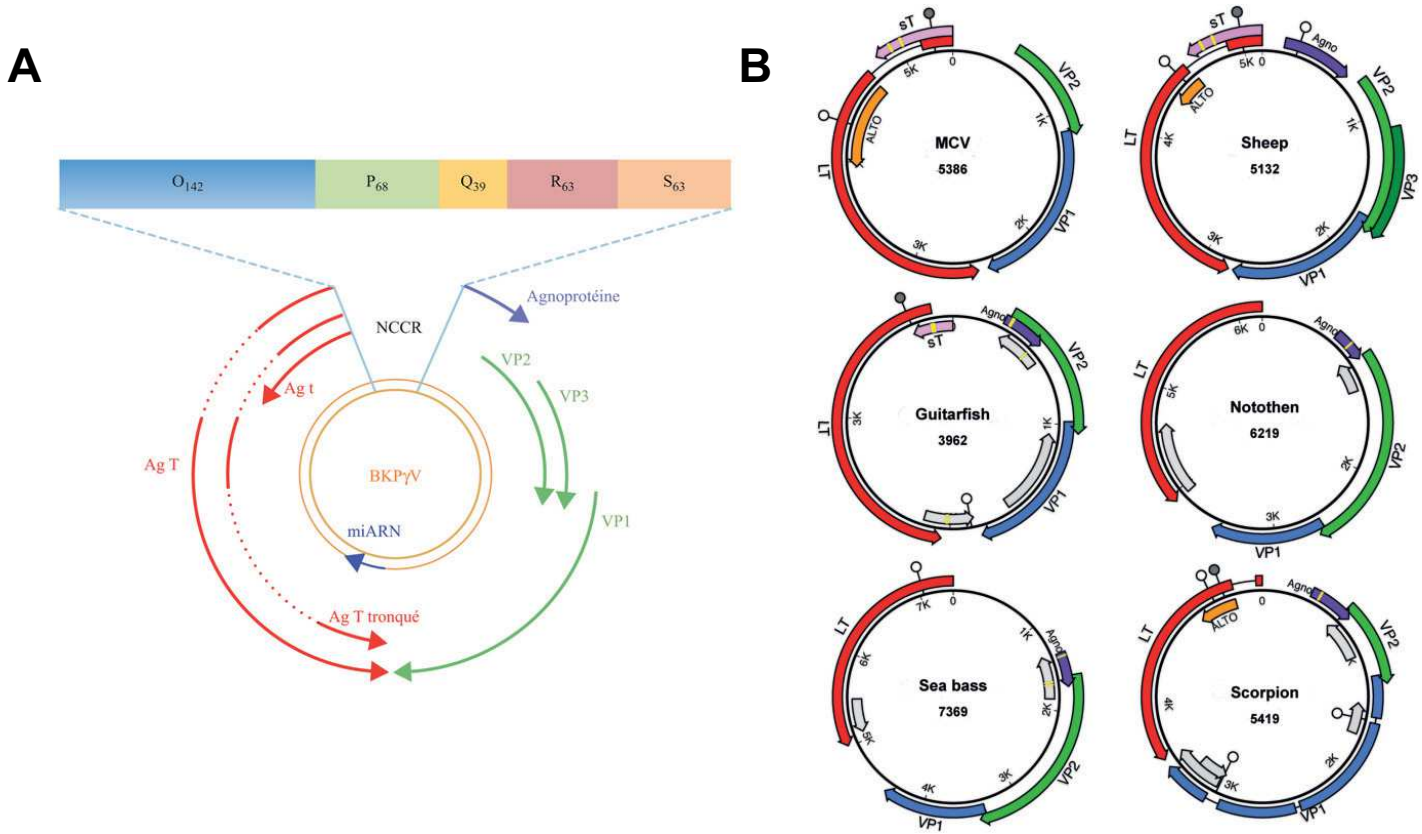

Figure 1.9 – **Représentation schématique A) du génome de BKPyV B) des génomes de six polyomavirus infectant différents hôtes d'origines taxonomiques diverses.** A) Le génome des BKPyV est organisé de la sorte : une région NCCR (un détail de cette région indique ses différentes séquences ainsi que leur taille chez un polyomavirus *wild-type*), une région, définie par la couleur rouge, comprenant les gènes précoces *LT* (indiqué comme Ag T dans la Figure), *ST* (indiqué comme Ag t) et *LT* tronqué (indiqué comme Ag T tronqué) et une troisième région, définie par les couleurs bleu et vert, comprenant les gènes tardifs *VP1*, *VP2* et *VP3*. B) Le détail de six génomes de polyomavirus infectant la souris (MCV), le mouton (Sheep), le poisson-guitare (Guitarfish), *Pagothenia borchgrevinki* (nothothen), le loup de mer (seabass) ainsi que le scorpion (scorpion) décrit une diversité dans l'organisation du génome, de ses gènes et dans la présence-absence de ces derniers. Certains génomes, comme le MCV ou le polyomavirus du mouton possèdent à titre d'exemple le gène *ALTO* (en orange sur cette figure), au contraire des autres polyomavirus. Au sein de cette Figure, la taille des génomes est donnée au centre de chaque schéma. Les gènes tardifs sont donnés par les couleurs vert, violet et bleu et les gènes précoces par les couleurs rouge et rose. Ces figures sont issues des articles de *Mazalrey et al.* et de *Buck et al.* [171, 173].

virus, bien qu'il l'améliore [189]. Ce gène semble aussi avoir une fonction oncogénique au sein des cellules infectées : il possède une affinité avec la protéine de l'hôte PP2A, elle même impliquée dans différents systèmes de protéines à l'origine des mécanismes du cycle et de l'apoptose cellulaires [189]. En dehors de cette interaction, le gène *ST* semble aussi avoir une interaction avec la voie de signalisation akt-mTOR, et modifierait la prolifération virale par ce biais [189].

À l'instar des gènes précoces, les gènes tardifs découlent tous d'un même ARNm. Ces gènes s'expriment de manière concomitante ou peu de temps après le début de la réplication virale [185]. Parmi ces gènes, on retrouve chez tous les polyomavirus les gènes des protéines virales *VP1* et *VP2* (pour *Viral Protein*). Un troisième gène viral *VP3* est retrouvé chez les BKPyV, JCPyV et MCPyV [171, 172]. Le rôle des protéines issues de ces gènes est explicite : elles font partie de la capside virale, structure icosaédrique entourant, protégeant et permettant le transport du matériel génétique du virus. Les deux (ou trois) protéines virales forment des pentamères qui, en s'emboîtant à l'intérieur du noyau de la cellule infectée, formeront une capside où VP1 est la seule protéine externe [190]. Les autres protéines virales vont, à l'instar de la charpente d'un toit, soutenir la structure de la capside et permettre la liaison de tous les pentamères qui la composent. Il a été établi que l'encapsidation des virions n'est pas régulée par un quelconque mécanisme spécifique, mais plutôt par l'accumulation du matériel génétique et protéique au sein de la cellule, et se ferait par l'interaction des protéines virales et des histones de l'hôte (à l'exception de H1) [191]. La protéine *VP1* représente donc la partie externe de la capside, et peut être considérée comme le point central des mécanismes d'interaction avec l'hôte [190]. Ce gène est donc soumis à une pression de sélection plus forte que les autres *VP*, et l'on observe une forte diversité intra-souche des *VP1* qui peut être due à des mécanismes d'échappement de l'immunité de l'hôte [171, 172]. Il semblerait que le rôle des protéines VP2 et VP3, en plus de participer à la structure de la capside, soit de stabiliser le maintien de la capside et surtout de permettre une entrée du virion au sein du noyau de la cellule [192, 193]. Certains polyomavirus possèdent d'autres gènes tardifs, comme celui de l'agnoprotéine chez les BKPyV et JCPyV [172]. Le rôle de l'agnoprotéine commence à être connu, oscillant entre régulation de l'expression des gènes tardifs et aide à l'assemblage de la capside lors de l'étape d'exocytose des virions, elle irait même jusqu'à modifier la prolifération virale pour permettre une infection persistante chez l'hôte [194–196]. Pour finir, l'agnoprotéine semble nécessaire au maintien d'une infection chez le JCPyV ; son absence conduit à un assemblage aberrant de la capside [194].

Certains polyomavirus comme les MCPyV, BKPyV, JCPyV et SV40 possèdent aux abords de la région tardive une courte région correspondant à un micro-ARN dont on suppose qu'il joue un rôle dans l'inhibition de l'expression des gènes précoces à la fin du cycle infectieux [197].

La NCCR est une région non-codante où se retrouvent les promoteurs des différents ARNm du génome, ainsi que l'origine de réplication (ou *ori*). Cette région est située entre le premier codon de la région précoce et le premier codon de la région tardive (Figure 1.9). La région NCCR peut être découpée en cinq régions distinctes O, P, Q, R et S [171]. La région O contient l'origine de réplication, mais aussi les régions de régulation et d'initiation de l'expression des gènes précoces [171]. Il s'agit d'une région hautement conservée, au contraire des régions P, Q, R et S. Le bloc P, Q, R constitue une séquence accueillant les facteurs de régulation de la transcription

de la cellule hôte, et peut donc être considéré comme une région modulatrice de l'expression des gènes viraux [171]. Pour finir, la région S possède le site promoteur de la transcription des gènes tardifs [171]. La région NCCR peut être modifiée par des substitution ou indels (pour INsertion-DÉLétion). La modification des régions O, P, Q, R et S, que ce soit par la modification, duplication, délétion ou l'inactivation de certaines régions, peut modifier certaines aspects du cycle de reproduction du virus [171].

Chez le BKPyV, malgré une diversité des régions NCCR, on considère qu'il existe à l'état naturel une forme « *wild-type* » (nommée NCCRww). Or, la modification de la pression de sélection due à l'atténuation du système immunitaire chez les patients immunosupprimés peut engendrer l'apparition de NCCR réarrangées (NCCRrr) pouvant posséder différents phénotypes décrivant la modification de l'expression des gènes précoces, tardifs ou des deux. Dans tous les cas, les mutants NCCRrr prolifèrent de manière significative par rapport aux souches NCCRww [171]. De ce fait, l'évolution et la diversité de la région NCCR chez un patient receveur d'une greffe de rein doit être analysée pour mieux comprendre l'évolution intra-hôte d'une infection par un BKPyV.

1.5.3 Classification et histoire évolutive des polyomavirus

1.5.3.1 Phylogénie des polyomavirus

La phylogénie des polyomavirus a été construite principalement à partir de l'analyse des gènes *LT* et *VP1*, et parfois du gène *VP2*. De par leur grande diversité dans l'organisation de leur génome et dans la présence et absence de certains gènes, mais aussi de la divergence entre les différents polyomavirus, il n'est pas possible d'effectuer une analyse sur leurs génomes complets. Sur la base de l'analyse du gène *LT*, deux classifications ont été proposées :

- La première, basée sur l'analyse du gène LT, est proposée par l'ICTV [161] et sépare les polyomavirus en quatre principaux clades à l'échelle du genre :
  - Le clade des *Alphapolyomavirus* qui contiendrait, plus de quarante souches virales infectant les mammifères dont les polyomavirus MCPyV, NJPyV ou encore le HPyV9. Ce clade serait, le plus important des polyomavirus.
  - Le clade des *Betapolyomavirus*, qui contiendrait environ trente souches virales infectant les mammifères dont les polyomavirus humains BKPyV, JCPyV, WUPyV, KIPyV ou encore le SV40.
  - Le clade des *Gammapolyomavirus*, où les polyomavirus infectent uniquement des hôtes aviaires.
  - Le clade des *Deltapolyomavirus* comprend uniquement les polyomavirus humains HPyV6, HPyV7, TSPyV et HPyV10 .
  - Un cinquième groupe rassemblant des polyomavirus de mammifères et de poissons sans préférence particulière.
- La deuxième, dont la version finale a été elle aussi proposée sur la base du gène *LT* par *Buck et al.*, résout la phylogénie des polyomavirus selon cinq clades distincts à l'échelle de la famille (Figure 1.10) [173] :

— les *Orthopolyomavirus*, qui infectent les mammifères et chez qui on retrouve les polyomavirus SV40, BKPyV, JCPyV, WUPyV, KIPyV et SA12.

— Les *Almipolyomavirus*, qui infectent eux aussi les mammifères et chez qui l'on retrouve la majorité des polyomavirus humains (HPyV6, HPyV7, HPyV9, , le polyomavirus STL, ...). Selon *Buck et al.*, ce clade peut être séparé en groupes comprenant les *Almipolyomavirus* et un groupe dénommé *Monominor* [173].

— Les *Avipolyomavirus*, qui constituent le clade regroupant les polyomavirus aviaires tels que le APyV. Au sein de ce clade, on retrouve un polyomavirus infectant les marsupiaux bandicoot, ce qui pourrait signaler des événements de recombinaison ou d'événements de changement d'hôte au cours de l'histoire évolutive des polyomavirus de ce clade.

— Les *Fish-polyomavirus*, un clade frère des Avipolyomavirus infectant les poissons.

— Les *Arthropod-polyomavirus*, infectant les arthropodes.

Dans le cadre d'une analyse sur le gène *VP1* (ou tout simplement sur les gènes tardifs), la phylogénie des polyomavirus est radicalement différente de celle proposée au sein de la Figure 1.10 (Figure 4 de *Buck et al.*). En effet, sous cette phylogénie, certains polyomavirus des clades *Orthopolyomavirus* et *Almipolyomavirus* forment un clade frère des polyomavirus de poissons, qui était connu avant sa révision sous le nom de *Wukipolyomavirus*. [198]. Par ailleurs, les *Avipolyomavirus* s'insèrent au sein du clade des *Orthopolyomavirus*, ce qui diffère de la classification de *Buck et al.* décrite ci-dessus. Ces incohérences entres les phylogénies des gènes précoces et tardifs pose un véritable problème dans la construction des phylogénies de polyomavirus. Il a été proposé par *Buck et al.* que plusieurs événements de recombinaison sont survenus lors de l'évolution des polyomavirus, et que ceux-ci expliquent une telle disparité entre les deux phylogénies, où celle des gènes tardifs est marquée par les dits événements de recombinaison (Figure 5 de *Buck et al.*).

Toujours d'après un modèle de *Buck et al.*, les polyomavirus coévoluent avec leurs hôtes depuis environ 500 millions d'années. Cette coévolution n'est pas stricte, mais pourrait être définie par une coévolution intra-hôte avec des changements d'hôte au cours de l'évolution (Figure 6 de *Buck et al.*). Au sein de ce modèle, les polyomavirus se diversifieraient au sein d'une espèce hôte, puis deviendraient parasites de nouveaux hôtes. Un tel modèle expliquerait les profils de coévolution retrouvés entre les *Orthopolyomavirus* et les *Almipolyomavirus*, où les polyomavirus appartenant à ces différents clades infectent des hôtes similaires [173].

1.5.3.2 Phylogénie des BKPyV

La phylogénie des BKPyV se structure autour de quatre grands génotypes I, II, III et IV [199, 200]. Au sein des génotypes I et IV, il existe un autre niveau taxonomique où coexistent quatre sous génotypes de I (I-a, I-b1, I-b2 et I-c) et six de IV (IV-a1, IV-a2, IV-b1, IV-b2, IV-c1 et IV-c2) [199–201] (Figure 1.11). Les quatres génotypes ont une résolution et une distance phylogénétique qui les séparent distinctement, à l'exception des génotypes II et III qui forment tous deux un clade suffisamment proche pour qu'on les considère comme des génotypes au même niveau que les sous-groupes des virus BK I et BK IV [199]. Les BKPyV I sont les plus

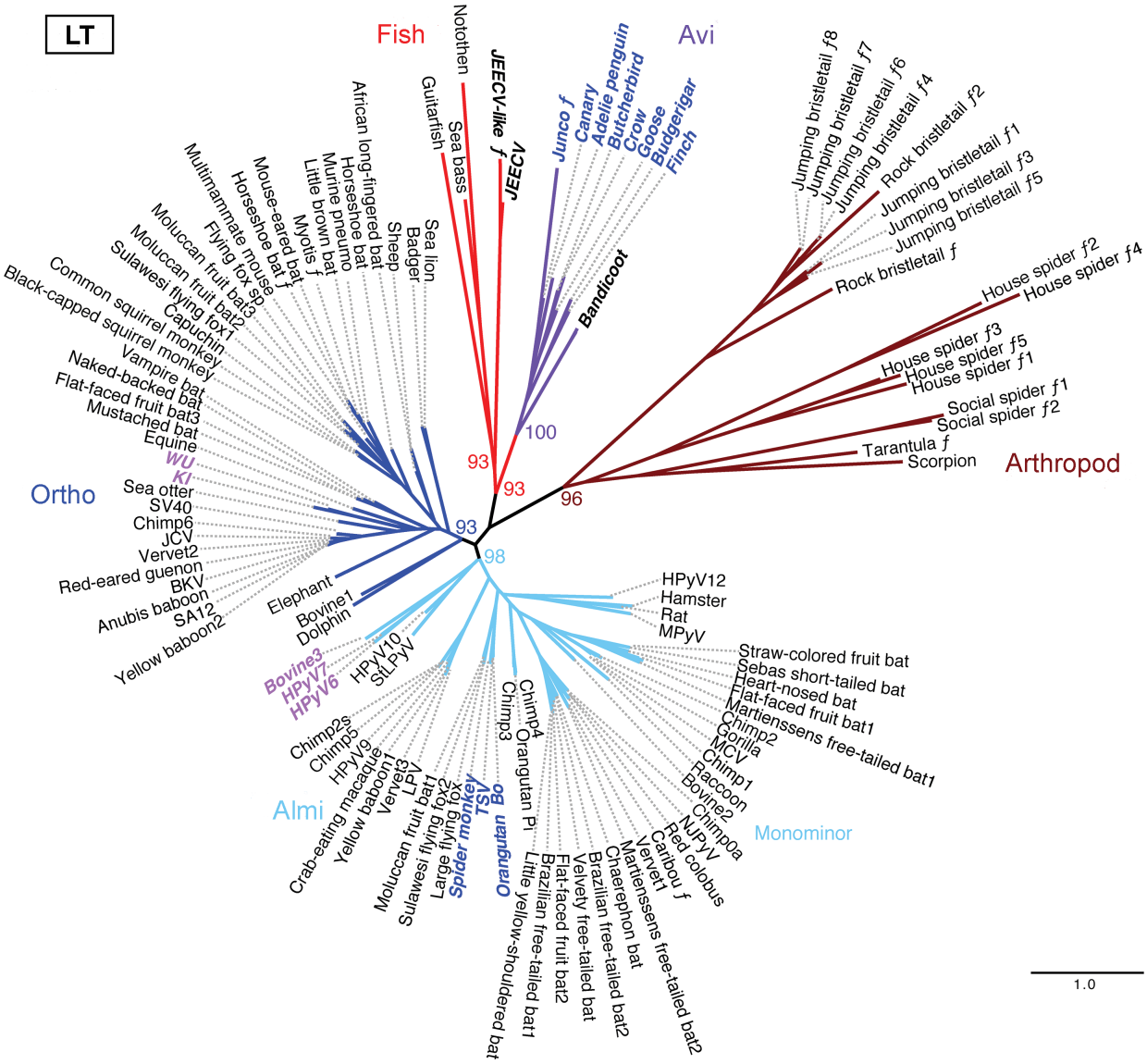

Figure 1.10 – **Phylogénies des polyomavirus connus à ce jour.** Cette phylogénie a été construite en Maximum de Vraisemblance à partir du gène *LT*. Cinq grands clades peuvent être délimités : les *Orthopolyomavirus* (en bleu, sur la gauche du graphe), les *Almipolyomavirus* (en cyan), les *Avipolyomavirus* (en bleu, sur la droite du graphe), les *Fish-polyomavirus* (en rouge) et les *Arthropod-polyomavirus* (en marron). Cette Figure est issue de l'article de *Buck et al.* [173]

représentés dans la population humaine (environ 80 % des BKPyV), suivis par les BKPyV IV (15 %, avec une prévalence plus élevée en Asie), puis par les BKPyV II et III qui ne représentent qu'une infime partie de la prévalence des BKPyV (environ 5 %) [199, 202].

1.5.4 Polyomavirus, infections humaines et aspects cliniques

1.5.4.1 Histoire infectieuse

Le point d'entrée d'une infection par un polyomavirus est souvent difficile à identifier, car une fois installé au sein de son hôte, celui-ci persiste dans l'organisme au sein de plusieurs tissus. Les tissus constituant le point d'entrée de l'infection sont donc souvent confondus avec les tissus réservoirs assurant la réplication virale, et ce aussi bien chez les patients immunosupprimés que sains. Selon le polyomavirus, le point d'entrée peut différer :

— Le BKPyV semble infecter leur hôte par le biais des voies respiratoires. Après une primo-infection, celui-ci est principalement retrouvé dans le rein, les glandes salivaires, les amygdales, le tractus respiratoire et les poumons. Bien que plus rare, certains cas de transmission verticale ont été observés entre la mère et l'enfant. Il a été aussi suggéré que ce virus se transmet par voie oro-fécale [172, 190].

— Bien que retrouvé dans les tissus rénaux, hématopoïetiques et lymphatiques, il est estimé que le point d'entrée des JCPyV sont les tissus de l'amygdale, car on les retrouve systématiquement dans ce tissu, et ils possèdent une capacité d'infection accrue lors d'un contact avec des tissus amygdaliens *in vitro*. Par ailleurs, ces virus pourraient eux-aussi être transmis par vois oro-fécale [190, 203].

— Les polyomavirus WUPyV et KIPyV, que l'on retrouve principalement dans les poumons, dans les tissus lymphoïdes nasaux et dans le système urinaire auraient pour point d'entrée les voies respiratoires [190, 204].

— Le MCPyV primo-infecte, en toute logique, leur hôte par un contact avec la peau. Les autres mécanismes de primo infection du MCPyV, s'ils existent, sont encore méconnus, mais l'on suppose une transmission par voies respiratoires ou par contact oro-fécal [190].

Le ratio de séro-prévalence, bien qu'il varie selon le polyomavirus, demeure élevé chez les enfants de moins de 10 ans : il est estimé que 65 à 90 % des jeunes individus ont connu un épisode infectieux avec le BKPyV alors que la séroprévalence des MCPyV et JCPyV est respectivement de 45 % et 50 % pour un même âge [205, 206]. Il est à noter que selon une étude américaine, quasiment tous les individus auraient connu un épisode infectieux par le BKPyV après l'âge de 10 ans [206, 207]. De ce fait, il n'est pas rare qu'un individu soit infecté par plusieurs polyomavirus sans, encore une fois, déclarer un seul symptôme [208] (Tableau 1 *Kean et al.*). Une fois le pic de charge virale atteint, la population des polyomavirus diminue au sein de leur hôte, jusqu'au moment où ils ne persistent plus qu'à bas bruit au sein des tissus réservoirs [171].

1.5.4.2 Maladies associées au BKPyV

La pathogénicité du BKPyV concerne généralement les individus immunosupprimés ayant reçu une greffe de reins ou de cellules souches hématopoïetiques [210–212]. À ce jour, de nom-

breux cas de cystites hémorragiques, de sténoses urétrales, d'inflammation du rein ou, dans une moindre mesure, de pneumopathies ou d'encéphalopathies dues au BKPyV ont été répertoriés [171]. Dans la majeure partie des cas, on peut considérer que les maladies associées au BKPyV, mais aussi à tous les autres polyomavirus pathogènes, sont dues à la réactivation des populations virales suite à la perte de l'immunocompétence de l'individu hôte [213].

Le BKPyV a été isolé pour la première fois chez un patient immunosupprimé souffrant d'une sténose urétrale [168]. La sténose urétrale est définie par une difficulté dans la miction, elle-même due à un rétrécissement des canaux urinaires. Après cet épisode marquant la découverte du BKPyV, plusieurs cas de sténose urétrale suivant une greffe de rein ou de cellules souches hématopoïetiques ont été répertoriés [168, 211, 214]. La détection d'une sténose urétrale par le BKPyV est simplement faite par la recherche du virus au sein de l'urine du patient [168, 211, 214].

La cystite hémorragique est définie par un saignement diffus de la muqueuse vésicale, provoquant des douleurs au niveau du pubis et des difficultés lors de la miction (qui sera bien souvent marquée par une hématurie). Dans les cas les plus sévères, la cystite hémorragique peut conduire à l'obstruction du système urinaire, à l'apparition d'hémorragies et finalement à une insuffisance rénale dans les cas les plus graves [212, 215]. Au même titre que les adénovirus ou encore le virus de la grippe A, le BKPyV est l'un des agents responsable de la cystite hémorragique, mais se démarque par une pathogénie exclusivement associée à une greffe de cellules souches hématopoïetiques [212, 216]. Il est estimé que 5 à 25 % des patients ayant reçu une greffe de cellules hématopoïetiques souffriront d'une cystite hémorragique due au BKPyV dans les mois suivants le début de la rémission [212, 217]. Les causes d'une cystite hémorragique associée au BKPyV sont encore méconnues, mais on estime que la réactivation du virus, associée à une dégradation des parois urothéliales par les différentes conséquences des traitements pré-greffe, de la greffe et post-greffe participent grandement à la mise en place de cette maladie [212]. Le diagnostic d'une infection par un BKPyV lors d'une cystite hémorragique est simple : il est basé sur la concentration du nombre de génomes viraux de BKPyV au sein d'un échantillon d'urine (*i.e.* la virurie observée) et de sang (*i.e.* la virémie observée). Les patients présentant une forte virurie (et qui possèdent bien souvent une virémie) ont une probabilité accrue de développer une cystite hémorragique [218, 219].

Le BKPyV est avant tout responsable de la PVAN (pour Polyomavirus Assiocated Nephropathy), une maladie se déclarant chez les patients receveur d'une greffe pendant les deux années suivant celle-ci. D'après plusieurs études, les polyomavirus provoquant la PVAN proviendraient du greffon du donneur, et non pas de l'individu receveur de la greffe [1, 202, 220]. La PVAN est aujourd'hui caractérisée par une augmentation soudaine de la charge virale des tissus épithéliaux des tubules du rein et du système urinaire [1, 171, 210]. Cette prolifération virale soudaine et incontrôlée conduit à la lyse d'un grand nombre de cellules et la dégradation des cellules épithéliales du rein conduit à l'infiltration des virions dans le système sanguin. La PVAN débute par une infiltration tissulaire de cellules inflammatoires, conduisant alors à une fibrose interstitielle et une atrophie tubulaire (dégradation du rein et de sa fonction). Au fur et à mesure de l'évolution de la maladie, le rein perd sa fonction, et dans les cas les plus graves, cette progression peut

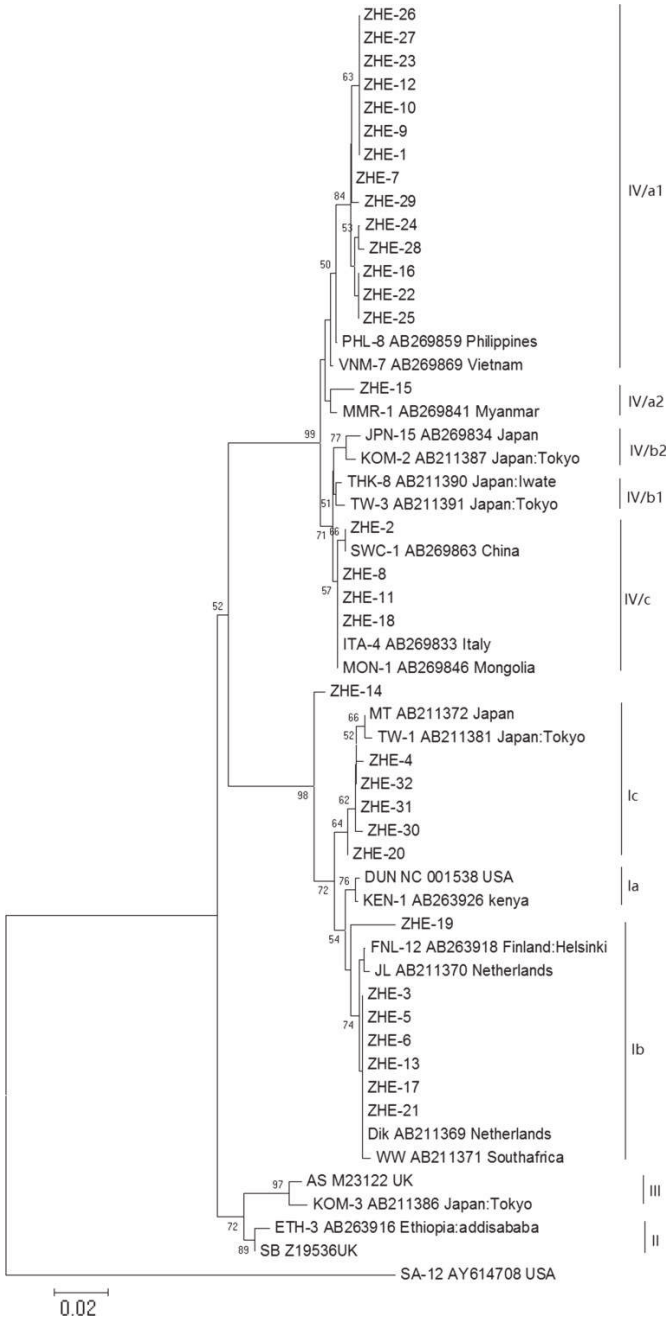

Figure 1.11 – **Phylogénie des BKPyV** Cette phylogénie a été réalisée en Maximum de Vraisemblance à partir de l'analyse des génomes de BKPyV. Seules les valeurs de « bootstrap » supérieures à 50 sont indiquées au sein de ce graphe. L'individu SA12, externe aux polyomavirus, sert d'« outgroup » et oriente la phylogénie. Les BKPyV sont répartis en quatre clades distincts (I, II, III et IV), et les clades I et IV se séparent à leur tours en sous-groupes. Cette phylogénie est issue d'un article de *Hu et al.* [201]

être irréversible et peut mener à un dysfonctionnement total du rein et donc à la perte du greffon [171]. On considère qu'environ 1 à 10 % des patients receveurs d'une greffe de reins souffriront d'une PVAN, et que 10 à 80 % de ces cas développeront une forme grave et irréversible de cette maladie [210].

Lors d'une PVAN, la réplication virale se situe primordialement au niveau des cellules épithéliales du tractus urinaire, même si une réplication virale soutenue peut être observée au niveau des cellules épithéliales des tubules du rein. Généralement, on considère un risque de PVAN lorsqu'un niveau de virémie supérieur à 10^4 copies ADN/mL est observé, car cette charge virale sanguine indique une forte réplication au niveau du tractus urinaire [1, 210]. Mais il est à noter qu'une forte réplication virale dans les tissus du tractus urinaire n'implique pas forcément l'apparition d'une PVAN. Il a été estimé que la majorité des patients développant une forte virurie ne développaient pas de virémie ni de PVAN [1, 221, 222]. Une fois le seuil de virémie de 10^4 atteint, il est suggéré que l'immunocompétence du patient doit être réactivée pour qu'il combatte l'infection virale [1]. En effet, il n'existe pas de traitement antiviral efficace à ce jour, et le seul traitement possible consiste en une modulation de l'immunosuppression du patient, mais ce type de traitement reste avant tout patient-dépendant, et peut conduire à un rejet du greffon par le système immunitaire [210].

1.6 Objectifs de la thèse

Cette thèse s'organise selon deux axes principaux dont le lien est le CUB.

Dans un premier temps, nous avons pour objectif d'améliorer notre compréhension du CUB et de son rôle chez les Vertébrés. Pour cela, nous développons tout d'abord COUSIN, un nouvel indice de mesure qui permet une analyse comparative du CUB d'une séquence requête par rapport à un jeu de données de référence, mais dont les résultats sont normalisés autour d'une hypothèse nulle décrivant un usage égal des codons synonymes. Ce nouvel indice est inséré au sein d'un programme éponyme permettant une analyse aisée et complète du CUB. L'indice et le programme COUSIN sont accessibles par le lien http://cousin.ird.fr. Par la suite, nous effectuons une analyse systématique du CUB et de l'histoire évolutive des gènes paralogues *PTBP*. Il existe chez les Vertébrés trois paralogues de ce gène appelés *PTBP1*, *PTBP2* et *PTBP3* [223]. *PTBP1* est fortement exprimé dans la majorité des tissus humains et est enrichi en GC, alors que les deux autres versions de ce gène sont enrichis en AT et présentent une expression spécifique à certains tissus [224]. Selon une expérimentation de *Robinson et al.*, le CUB aurait un effet sur la tissu-spécificité des *PTBP* [224]. Nous proposons d'explorer l'histoire évolutive et le CUB des gènes *PTBP*, dans l'optique d'y déceler des signatures évolutives vers une expression tissu-spécifique.

La deuxième partie de ce projet s'insère au sein de l'ANR BK-NAB du laboratoire de virologie du CHU de Nantes (en partenariat avec les laboratoires MIVEGEC) et vise à proposer des solutions bioinformatiques et analytiques dans le cadre de l'observation longitudinale de la diversité et de l'évolution des souches de BKPyV chez des patients receveurs d'une greffe de rein. Les objectifs de cet ANR sont de mieux comprendre l'évolution virale du BKPyV chez un patient après modulation du système immunitaire, car il est estimé que sans pression de la part

du système immunitaire, une véritable diversité intra-hôte des BKPyV se met en place. Cette diversité, qui ne se retrouve pas à l'état naturel, peut notamment modifier :

— La composition nucléotidique de *VP1*, provoquant une diversité au niveau des épitopes de la capside. Celle-ci peut freiner l'efficacité du système immunitaire si celui-ci est remis en place pour lutter contre la prolifération virale. En effet, une diversité au niveau des épitopes de la capside des BKPyV permettrait aux virions d'échapper au système immunitaire de l'hôte [171, 172].

— L'organisation et la structure de la région NCCR. Comme nous l'avons vu précédemment, les modifications de la région NCCR peuvent conduire à l'apparition de populations virales avec une forte expression des gènes précoces, tardifs ou de tous les gènes. Ces populations, qualifiées de mutantes, pourraient remplacer les naturelles, et aggraver la prolifération des BKPyV au sein de l'hôte receveur de la greffe [171, 172].

Pour observer l'évolution des populations et de la diversité virale au cours d'une infection, l'ANR BK-NAB propose de prélever de multiples échantillons de sang et d'urine sur des patients au cours de la rémission de leur greffe. Chaque aspect clinique du patient (âge, sexe, évolution de la rémission, apparition ou disparition d'une PVAN, virurie et virémie des BKPyV, ...) est enregistré, et le matériel viral est séquencé selon un méthode de séquençage PacBio CCS. L'observation de la diversité des génomes, couplée aux données de virémie, de virurie et du patient, permettrait alors de déduire si l'apparition de nouveaux génotypes serait à l'origine d'une évolution de l'infection vers une PVAN. Nous avons développé un nombre d'outils et d'éléments de recherche pour analyser de telles données. Nous proposons par ailleurs une analyse de l'évolution des polyomavirus humains et de leur CUB, dans l'optique d'observer des motifs particuliers qui seraient apparus au cours de l'évolution de ces entités.

Partie Un

I

ÉTUDE DU CUB ; NOUVELLES APPROCHES MATHÉMATIQUES, INFORMATIQUES ET ANALYTIQUES.

Chapitre Deux

2

COUSIN, UNE APPROCHE NORMALISÉE DE LA MESURE DU CUB

Au sein de ce premier chapitre, nous présentons COUSIN (pour *COdon Usage Similarity Index*), un indice de mesure comparatif du CUB novateur. Notre désir de créer un nouvel indice puise son origine dans la volonté de proposer une solution qui, *à contrario* de l'indice CAI, permettrait aussi bien de comparer le CUB d'une requête par rapport à une référence qu'à un usage équiprobable des codons. Cet indice a notamment été développé dans une optique de comparaison du CUB d'un virus à celui de son hôte. Nous avons par la suite développé le programme `COUSIN`, qui permet la détermination du CUB de séquences requêtes à l'aide de l'indice COUSIN (entre autres) et qui propose un panel de tâches permettant d'approfondir l'analyse du CUB. COUSIN a fait l'objet d'une publication scientifique au sein de la revue scientifique d'Oxford GBE (PMID : 31800035) [94].

2.1 Indice COUSIN

L'indice COUSIN compare les fréquences des codons synonymes d'une séquence requête à celles d'un jeu de données de référence. Les résultats de cette comparaison sont ensuite normalisés autour d'une Hypothèse Nulle H_0 décrivant un usage équiprobable des codons. Les notations utilisées pour décrire les étapes du calcul de COUSIN sont indiquées dans le tableau 2.1.

Selon certains auteurs, la composition en acides aminés d'une protéine peut avoir un impact sur le biais d'usage des codons du CDS associé [148]. Nous avons donc conceptualisé deux versions de notre index : $COUSIN_{18}$, où les 18 familles de codons synonymes participent de manière équivalente au score final et $COUSIN_{59}$, où chaque famille contribue proportionnellement à la fréquence de l'acide aminé au sein de la requête. La comparaison des scores 18 et 59 permet de déterminer l'impact de la composition en acides aminés sur le CUB.

Tout d'abord, un premier score de déviation est calculé à partir des fréquences au sein de l'H_0 et de la référence :

$$\mathrm{dev}_{\mathrm{c,a}} = f_{c,a}^{\mathrm{ref}} - f_{c,a}^{H_0} \tag{2.1}$$

où $f_{c,a}^{\text{ref}}$ représente la fréquence du codon c de l'acide aminé a au sein de la référence, et $f_{c,a}^{H_0}$ celle sous l'Hypothèse Nulle. Ensuite, un score de pondération est calculé pour la référence,

$$\mathrm{W}_{\mathrm{c,a}}^{\text{ref}} = f_{c,a}^{\text{ref}} \times \text{dev}_{c,a} \tag{2.2}$$

et pour la requête,

$$\mathrm{W}_{\mathrm{c,a}}^{\text{que}} = f_{c,a}^{\text{que}} \times \text{dev}_{c,a} \tag{2.3}$$

où $f_{c,a}^{\text{que}}$ représente la fréquence du codon c de l'acide aminé a au sein de la requête. Le score $COUSIN_{18}$ d'un acide aminé provient du ratio de la somme des scores de poids de la requête par celle de la référence :

$$\mathrm{COUSIN}_{18}^{\mathrm{a}} = \frac{1}{\mathcal{N}} \times \frac{\sum\limits_{c \in k_a} \mathrm{W}_{c,a}^{\text{que}}}{\sum\limits_{c \in k_a} \mathrm{W}_{c,a}^{\text{ref}}} \tag{2.4}$$

où $\mathcal{N}$ représente le nombre qualitatif d'acides aminés en commun dans la requête et la référence et k_a le nombre de codons synonymes codant l'acide aminé a.

Pour $COUSIN_{59}$:

$$\mathrm{COUSIN}_{59}^{\mathrm{a}} = f_a^{\text{que}} \times \frac{\sum\limits_{c \in k_a} W_{c,a}^{\text{que}}}{\sum\limits_{c \in k_a} W_{c,a}^{\text{ref}}} \tag{2.5}$$

La différence entre les deux versions de cet indice réside dans le fait que dans le premier cas, nous considérons que tous les acides aminés sont représentés de manière équiprobable dans la

TABLE 2.1 – Notations utilisées pour définir les indices COUSIN

Symbole	Description
c	Codon
a	Acide aminé
f	Fréquence
ref	Référence
que	Requête
H_0	Hypothèse nulle
L	Taille de la requête
k_a	Ensemble des codons codant l'acide aminé a
$\mathcal{A}$	Ensemble des acides aminés présents dans la requête et la référence
$\mathcal{N}$	Nombre d'acides aminés présents dans la requête et la référence

séquence analysée ($1/\mathcal{N}$), alors que dans la deuxième version, nous ajoutons les fréquences des acides aminés dans le calcul du score COUSIN (f_a^{que}). Les scores $COUSIN_{18}$ et $COUSIN_{59}$ finaux sont obtenus en moyennant les scores COUSIN de chaque acide aminé présents dans la requête et dans la référence.

$$\text{COUSIN}_{18} = \sum_{a \in \mathcal{A}} \text{COUSIN}_{18}^{a} \tag{2.6}$$

$$\text{COUSIN}_{59} = \sum_{a \in \mathcal{A}} \text{COUSIN}_{59}^{a} \tag{2.7}$$

De par sa construction, un score COUSIN propose une détermination du CUB de la requête avec une forte significativité biologique (Figure 2.1) :

1. un score de 0 indique un CUB similaire à celui de l'H_0 (c'est-à-dire que la requête possède au mieux un faible CUB)
2. un score de 1 indique un CUB similaire à celui de la référence
3. un score entre 0 et 1 indique un CUB similaire à la référence mais plus atténué
4. un score supérieur à 1 indique un CUB similaire à la référence mais plus important.
5. un score inférieur à 0 indique que la séquence requête n'a pas le même CUB que la référence.
6. plus un score s'éloigne de 0, plus son CUB s'éloigne de l'H_0 et donc d'un usage équiprobable des codons synonymes.

Les limites prises par l'indice COUSIN dépendent du CUB de la référence : plus le CUB de la référence est proche de celui de l'H_0, plus grande est la gamme de résultats de COUSIN. À titre d'exemple, le faible CUB du génome d'*Homo sapiens* fait que les scores obtenus ont une variance élevé (valeurs extrêmes : [-4.48 ; 6.13]). Chez *Plasmodium falciparum*, qui possède un fort CUB global, les possibles scores COUSIN ont une variance moins forte (valeurs extrêmes : [0.15 ; 1.35]). Pour faciliter l'interprétation de COUSIN, l'outil présenté ci-dessous propose d'imposer des limites artificielles.

Pour permettre une comparaison aisée de COUSIN avec l'indice CAI, nous avons conceptualisé une nouvelle version, appelée CAI_{18}, qui supprime la diversité en acides aminés de l'équation [47]. Le CAI, dans sa version classique (voir Introduction), pourrait être ainsi appelé CAI_{59} car à l'instar du $COUSIN_{59}$, il prend en compte la composition en acides aminés dans son calcul. Le CAI_{18} diffère simplement dans le calcul final du score, où cette fois-ci l'occurrence de chaque codon est divisée par l'occurrence de l'acide aminé au sein de la séquence étudiée :

$$\text{CAI}_{18} = \left(\prod_{a \in \mathcal{A}} \prod_{c \in k_a} \frac{\text{Occ}_{c,a}^{(\text{que})}}{\text{Occ}_{a}^{(\text{que})}} \times w_{c,a} \right)^{\frac{1}{\mathcal{N}}} \tag{2.8}$$

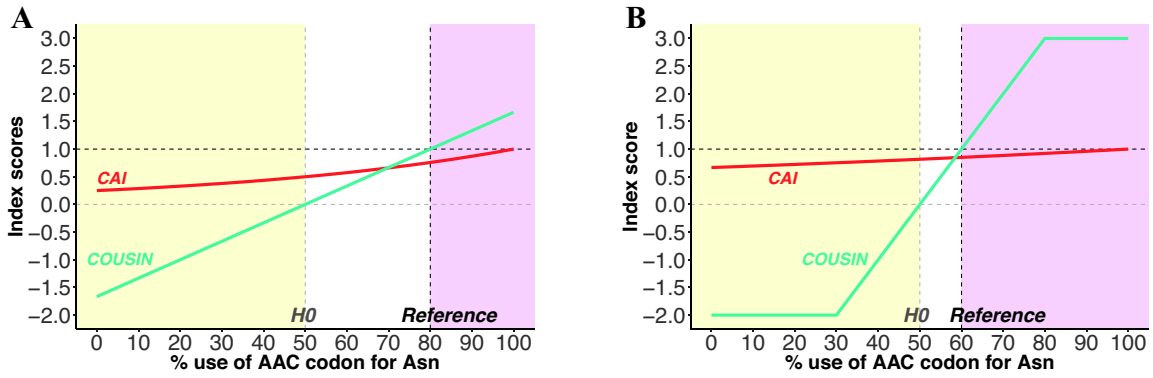

Figure 2.1 – **Scores COUSIN (courbe bleue) et CAI (courbe rouge) pour un jeu de données hypothétiques de séquences requêtes avec une différence dans la fréquence des codons synonymes AAC et AAU de l'asparagine (abscisse)**. Les valeurs sont calculées à partir de deux jeux de données hypothétiques possédant A) un fort CUB AAC :AAU de 80 :20 et B) un faible CUB AAC :AAU de 60 :40. Les lignes verticales indiquent un CUB nul (H_0, en gris) et celui de la référence (en noir). Les lignes horizontales décrivent les valeurs clefs prises par COUSIN qui correspondent à l'H_0 (en gris) et à la référence (en noir). La zone jaune indique les séquences requêtes avec un CUB opposé à celui de la référence, et la zone rose les requêtes avec un CUB similaire (et parfois exacerbé) à celui de la référence. De par leur définition, les valeurs COUSIN ont toujours une valeur de 0 et de 1 pour représenter un CUB similaire à l'H_0 et à la référence, et ce de manière indépendante au CUB de la référence. Par définition, le score CAI prend des valeurs allant de 0 à 1. Les scores COUSIN dépassant -3 et 4 sont omis pour faciliter la visualisation de ce graphe. Cette Figure est issue de l'article de *Bourret et al.* [94].

Où $\text{Occ}_a^{(\text{que})}$ est l'occurrence de l'acide aminé a au sein de la séquence requête, $\text{Occ}_{c,a}^{(\text{que})}$ est l'occurrence du codon c de l'acide aminé a au sein de la séquence requête et $w_{c,a}$ le score d'adaptativité relative du codon c de l'acide aminé a. Encore une fois, la comparaison des scores CAI_{18} et CAI_{59} permet de quantifier l'importance de la composition en acides aminés d'une séquence.

2.2 Programme `COUSIN`

Nous avons développé en parallèle de l'indice COUSIN un programme Python3 éponyme. Celui-ci détermine le CUB d'un jeu de séquences requêtes à l'aide de COUSIN et de huit autres indices. Plusieurs tâches annexes proposent à l'utilisateur d'approfondir l'analyse du CUB sous plusieurs angles. Pour faciliter l'utilisation de cet outil, un site web `COUSIN` a été déployé (cousin.ird.fr).

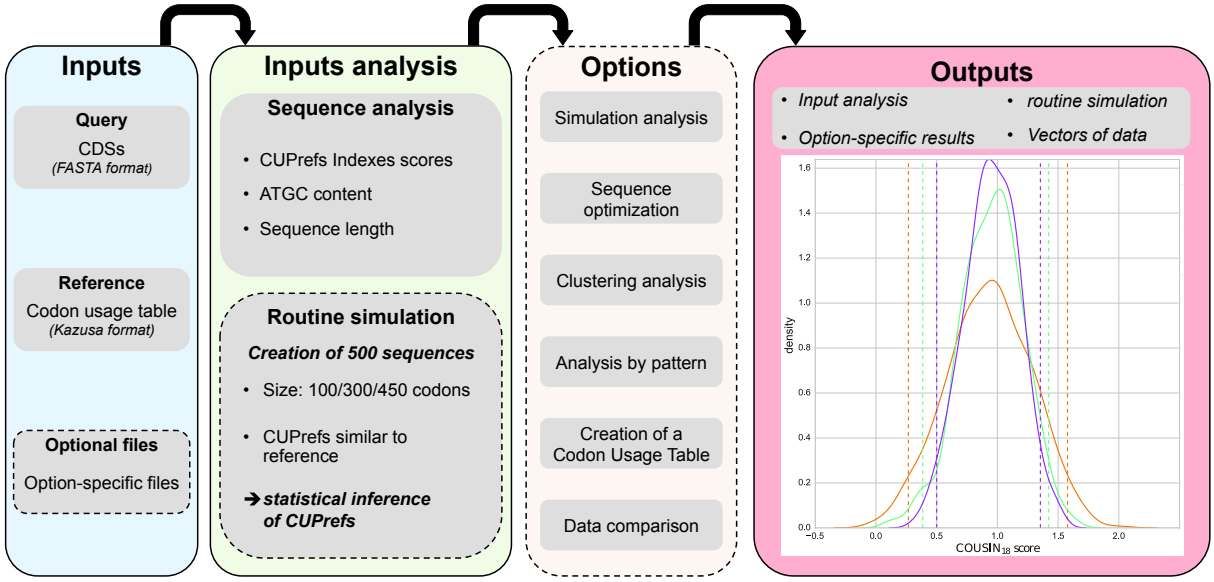

Figure 2.2 – **Architecture du programme COUSIN.** L'utilisateur insère en données d'entrée un fichier fasta contenant les séquences requêtes et un jeu de données de référence ainsi que des informations complémentaires si nécessaire. Le programme peut n'effectuer qu'une tâche de calcul basique du CUB et des spécificités en nucléotides des séquences requêtes, mais il est aussi possible de choisir parmi six tâches pour compléter et parfaire l'analyse du CUB. Différents fichiers sont rendus en sortie aux formats textuels et graphiques.

2.2.1 Architecture COUSIN

Un détail de l'architecture est donné en Figure 2.2. Le programme `COUSIN` repose sur une architecture en trois temps. Tout d'abord, l'utilisateur entre les données requêtes — des séquences au format fasta — la référence — au format CUT de *Kazusa* — et des informations propres à la nature du calcul demandé par l'utilisateur [133]. Ensuite, le CUB de chaque séquence est calculé, puis une étape d'approfondissement de la détermination du CUB est effectuée selon les demandes de l'utilisateur. Les résultats de l'analyse sont donnés sous des formes textuelles et graphiques.

2.2.2 Indices de calcul du CUB

Le programme permet de calculer le CUB d'une séquence requête à l'aide de neuf indices, mais détermine aussi deux scores concernant la composition en acides aminés de la requête :

— $COUSIN_{18}$/$COUSIN_{59}$
— CAI_{18}/CAI_{69} [47, 94]
— ENC [142]
— FOP [37]

— ICDI [134]
— CBI [38]
— SCUO [140, 146]
— MCB [139]
— χ^2 [137]
— AROMA [225]
— GRAVY [226]

Il s'agit à ce jour de l'outil le plus complet de calcul du CUB en permettant, à titre comparatif, de calculer le CUB à l'aide de neuf indices contre cinq avec le programme `CodonW` [150]). En plus de proposer les indices COUSIN, CAI et ENC, le programme détermine aussi le score d'indices qui n'ont jamais été implémentés au sein d'un outil accessible tels que le MCB [139] et le χ^2 [137].

2.2.3 Calcul du CUB

Après avoir reçu les données requêtes de l'utilisateur, COUSIN effectue automatiquement plusieurs calculs sur la nature des séquences. Celles-ci sont filtrées selon des critères définissant le caractère codant de la séquence : divisibilité par 3, appartenance stricte à l'alphabet de l'ARN ou de l'ADN et absence de codons STOP à d'autres endroits qu'à la fin de la séquence. Ensuite, la taille de la séquence et le contenu en GC aux trois bases des codons sont calculés. Le CUB des séquences est déterminé via les neuf différents indices cités ci-dessus. Pour finir, une étape de simulation est proposée à l'utilisateur (Figure 2.3). L'objectif de celle-ci est de simuler un jeu de données de 500 séquences requêtes avec une optimisation de leurs codons en *random-guided*. À partir de la distribution COUSIN de ces séquences simulées (ou de tout autre indice), des intervalles de confiance 95 % et 99 % sont comparés au score de la séquence requête, assurant ainsi la fiabilité statistique du score de l'indice concerné.

2.2.4 Fonctionnalités du programme COUSIN

Au-delà du calcul du CUB, `COUSIN` propose six fonctionnalités pour approfondir l'analyse du CUB des requêtes :

— Une option d'optimisation des gènes. qui modifie le CUB de la séquence pour qu'elle suive celui d'un autre jeu de données [157].
— Une simulation en *random-guided* à partir des séquences requêtes est effectuée i) sur le CUB ii) sur le CUB et la composition en acides aminés. Les intervalles de confiance de la gamme des scores COUSIN obtenus donnent alors des indications sur la significativité du CUB de la requête, notamment sur le rôle de la composition en acides aminés dans le CUB observé [152, 158].
— Une analyse d'agrégation des données selon le choix de l'utilisateur. À titre d'exemple, cette étape permet à l'utilisateur de rassembler des séquences par un motif dans leur *header* (nom de séquence) et de déterminer conjointement leur CUB.

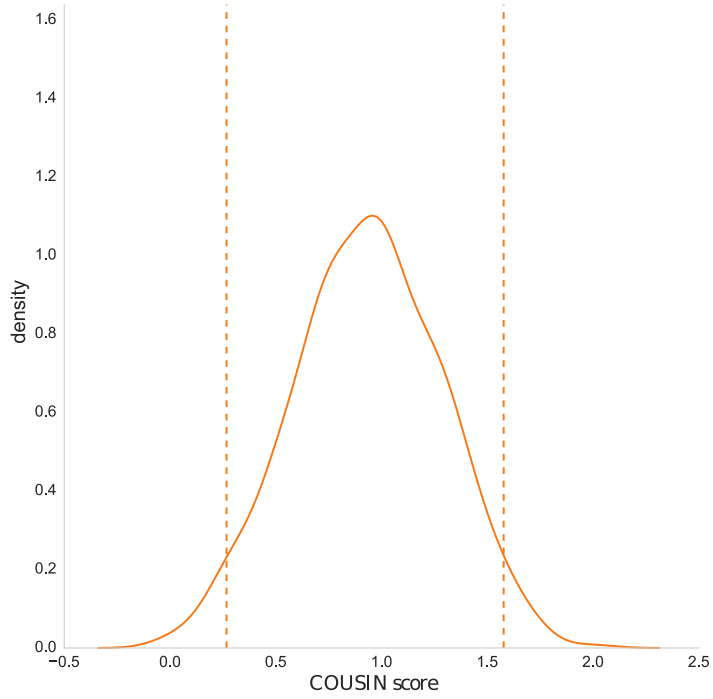

Figure 2.3 – **Courbes de densité des scores $COUSIN_{18}$ obtenus à la suite de l'étape de simulation de données**. Les courbes représentent des données simulées possédant une taille de 100 (orange), 300 (cyan) et 450 (violet) codons. Il s'agit des tailles moyennes retrouvées chez les petites protéines, les protéines de bactéries et d'eucaryotes. Les distributions en cloche obtenues permettent de définir les intervalles de confiance à 95 % et 99 % (représentés par les lignes verticales ayant la même couleur que les courbes associées) pour chaque ensemble de séquences simulées. Une simple comparaison de la séquence requête aux intervalles de confiance permet d'inférer statistiquement le CUB de la séquence requête par rapport à la référence.

— Une analyse d'agrégation des données selon une approche k-means ou de clustering hiérarchique. Les séquences requêtes sont groupées selon leurs CUB, leur contenu en GC ou encore selon les scores qu'elles ont obtenu pour de multiples indices. Une telle étude permet alors de déterminer efficacement les groupes de séquences requêtes ayant un CUB similaire.

— Un créateur de table d'usage des codons. À partir d'un jeu de données de séquence requêtes, il est possible de créer une CUT au format de *Kazusa* [133].

— Un comparateur de CUB. Ici, nous proposons d'estimer les distances euclidiennes du CUB entre deux jeux de données (jeu de séquences, tables d'usage des codons, valeurs RSCU, ...)

2.3 Analyse COUSIN

Pour démontrer la puissance de l'indice et du programme COUSIN, nous avons effectué une analyse comparative des scores COUSIN et CAI obtenus suite à l'analyse des CDS d'organismes appartenant à différents groupes du vivant.

2.3.1 Matériel et méthodes

Les génomes complets de huit organismes ont étés sélectionnés au sein de la base de données NCBI sous le format Genbank [227] : deux mammifères *H. sapiens* et la souris grise *Mus musculus*, la poule domestique *Gallus gallus*, l'arabette des dames *Arabidopsis thaliana*, le protiste *Plasmodium falciparum*, *S. cerevisiae* et deux bactéries *E. coli* et *Streptomyces coelicolor*. Grâce à la fonction `extractfeat` d'`EMBOSS`, nous avons extrait les CDS totaux de chaque organisme [154]. Nous avons ensuite conservé les CDS ayant une taille supérieure à 300 nucléotides. Pour supprimer une quelconque redondance au sein des gènes isoformes (*i.e* un gène pouvant donner différentes protéines), nous n'avons gardé que le premier représentant référencé dans les fichiers Genbank de chaque organisme. Les informations relatives aux numéros d'accès des génomes, au nombre de CDS extraits et à leur contenu en GC3 sont indiqués dans la Table 2.2. De plus amples informations sur le contenu en GC3 des CDS sont données au sein du Tableau Annexe B.1.

À partir des CDS extraits, nous avons construit des jeux de données de référence au format de *Kazusa* avec le programme `COUSIN`. Le CUB des références représente donc le CUB moyen des organismes. Pour finir, nous avons calculé les scores COUSIN et CAI de chaque CDS en le comparant à la référence de leur propre organisme. De ce fait, le CUB de chaque CDS est comparé à la moyenne de tous les CDS de l'organisme associé. Nous calculons le score de position Huber-M et les valeurs MAD (*Median Absolute Deviation* ou déviation absolue de la médiane en français) des distributions COUSIN et CAI observées pour chaque organisme [228].

2.3.2 Résultats

COUSIN représente avec exactitude et une forte capacité d'interprétation le CUB des organismes

La distribution du CUB des CDS de chaque organisme est donnée en Figure 2.4. Les courbes observées avec l'indice COUSIN sont toutes centrées autour de la valeur 1 (score qui représente la référence, et donc ici la moyenne du CUB de l'organisme). La valeur de l'estimateur Huber-M est elle aussi proche de 1 pour chacune des distributions. Il est à noter que malgré cette similarité dans le score COUSIN moyen, les distributions de certains organismes sont particulières et soulignent l'existence de disparités dans les CUB observés au sein du génome d'un organisme et entre les espèces. Chez *E. coli*, *A. thaliana* et *S. cerevisiae*, on observe une distribution normale des CDS. Chez *P. falciparum* et *S. coelicolor*, les distributions sont elles aussi normales, mais la variance y est beaucoup plus réduite. Une telle différence dans la distribution du CUB des CDS de ces organismes peut s'expliquer par les particularités des génomes de *P. falciparum* et *S. coelicolor*, qui tendent vers un biais de composition nucléotidique extrême. Les CDS de *P. falciparum*

TABLE 2.2 – **Récapitulatif des informations sur les CDSs complets des huit organismes inclus dans cette analyse.** Le tableau indique le nom de l'espèce, la référence du génome (issu de la base de données NCBI), le nombre de CDS gardés pour l'analyse (évalué par la suppression des isoformes et des CDS de petite taille), le nombre total de CDS extraits des fichiers genbank, le ratio entre les deux dernières variables ainsi que le contenu en GC3 des CDSs gardés pour la suite de l'analyse.

Espèce	**Référence**	**CDS sélectionnés**	**Nombre total de CDS**	**Ratio**	**Contenu en GC3**
Escherichia coli	K-12 substr. MG1655	3244	4319	0.8	54.9%
Streptomyces coelicolor	A3(2)	6356	8152	0.8	92.3%
Saccharomyces cerevisiae	S288C (assembly R64)	5549	5989	0.9	39.2%
Plasmodium falciparum	3D7 (assembly ASM276v1)	4773	5334	0.9	17.8%
Homo sapiens	Assembly GRCh38.p11	18492	115320	0.1	60.0%
Gallus gallus	Assembly GRCg6a	15751	49767	0.3	60.6%
Mus musculus	Assembly GRCm38.p6	20393	79262	0.3	58.6%
Arabidopsis thaliana	Assembly TAIR10	24774	48148	0.5	42.7%

sont riches en AT3 et ceux de *S. coelicolor* riches en GC3. De par cette composition nucléotidique particulière, les codons synonymes retrouvés au sein de ces deux organismes sont orientés vers les nucléotides surreprésentés avec, par exemple, une surutilisation du codon UUC chez *S. coelicolor* et du codon UUU chez *P. falciparum* pour coder la phénylalanine. Ainsi, les CDS de ces deux organismes possèdent un CUB similaire et orienté vers la composition nucléotidique des génomes respectifs, ce qui explique la faible diversité dans le CUB observé. Pour finir, les trois génomes de Vertébrés ont une distribution bimodale décrivant des groupes de CDS ayant des particularités au sein de leurs CUB. Ce dernier point sera discuté en détail dans la prochaine section pour *H. sapiens* et *G. gallus*.

Il est plus difficile de tirer des conclusions sur les observations des distributions des scores CAI. Chaque courbe possède une distribution normale mais présente des variances et moyennes différentes. Cela est principalement dû à la nature de ce que mesure le CAI. En effet, bien que souvent considéré comme un indice de mesure du CUB, il mesure l'« optimalité » des codons par rapport à sa référence. De ce fait, seules des informations relatives à cette optimisation peuvent être extraites de ces courbes, sans que l'on ait une quelconque idée du CUB de la requête ou de sa relation avec la référence. Par ailleurs, la valeur des scores CAI est directement impactée par le CUB retrouvé au sein de la référence, ce qui fait que les distributions obtenues sont propres à chaque organisme et ne peuvent pas être comparées (voir Figure 2.1 pour une exemplification de l'impact de la référence sur la distribution du CUB).

Nous n'observons pas de variation remarquable entre les versions 18 et 59 des indices COUSIN et CAI (Figure 2.5 A et D). Dans cette étude, il semblerait que le CUB des gènes soit faiblement impacté par la composition en acides aminés observée. La comparaison des scores COUSIN et CAI chez *H. sapiens* indique une forte corrélation entre ces deux indices, malgré l'incapacité flagrante de l'indice CAI de déceler clairement les différentes populations de CDS au sein de cet organisme (Figure 2.5 B et C). Une telle corrélation renforce l'idée de l'aspect pluridisciplinaire de COUSIN, où celui-ci serait aussi capable de déterminer, à une certaine mesure, l'« optimalité » d'une séquence par rapport à une référence.

Nous avons testé la corrélation des scores de CAI et COUSIN avec le contenu en GC3 des gènes pour chaque organisme (Figures Annexes B.1, B.2, B.3 et B.4) . Il apparaît que les scores COUSIN des gènes suivent plus fidèlement le contenu en GC3 que le score CAI, bien que les corrélations soient fortes et significatives dans chaque cas. Par ailleurs, la distribution du contenu en GC3 des gènes des Vertébrés montre encore une bimodalité similaire à celle de COUSIN. Cela indique la claire corrélation entre les deux indices et la capacité de COUSIN à faire ressurgir des spécificités propres à l'organisme.

COUSIN met en évidence les spécificités en CUB des vertébrés

Comme évoqué précédemment, les distributions des scores COUSIN sont non-unimodales chez *M. musculus* et affichent clairement une bimodalité chez *H. sapiens* et *G. gallus*. Dans cette section, nous allons explorer le CUB des gènes de *H. sapiens* et *G. gallus* en fonction de leur contenu en GC3, ou encore, de leur position intra-chromosomique chez *H. sapiens* et inter-chromosomique chez la poule (Figures 2.6 et 2.7). Il est établi que les chromosomes humains sont agencés en isochores, c'est-à-dire en région intercalaires riches en GC ou en AT, et que

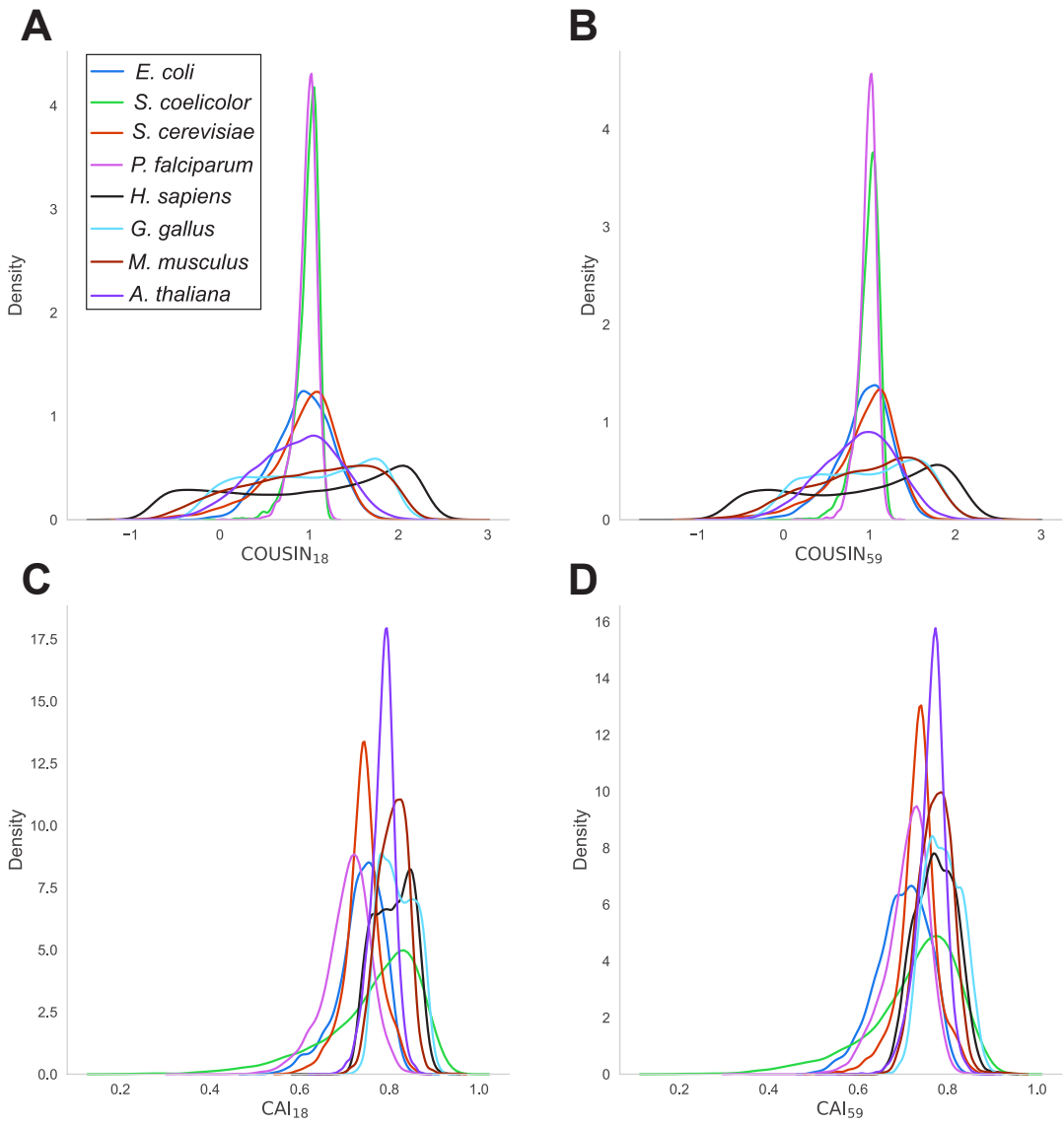

Figure 2.4 – **Courbes de densité des scores $COUSIN_{18}$ (A), $COUSIN_{59}$ (B), CAI_{18} (C) et CAI_{59} (D) pour chacun des organismes étudiés (voir légende sur la Figure).** Pour chaque CDS, les valeurs COUSIN et CAI ont été estimées à partir d'un jeu de données de référence représentant le CUB moyen de l'organisme. La normalisation de COUSIN centre les distributions aux alentours d'une valeur Huber-M de 1, permettant une rapide identification de la dispersion des CDS au sein de chaque organisme, et d'en déduire la diversité en CUB de chaque organisme par rapport à la référence et à l'H_0 (*e.g.*, les courbes à petite variance de *S. coelicolor* en vert). Au sein des résultats COUSIN, on observe une bimodalité de la distribution pour *H. sapiens* et *G. gallus*. Cette Figure est issue de l'article de *Bourret et al.* [94].

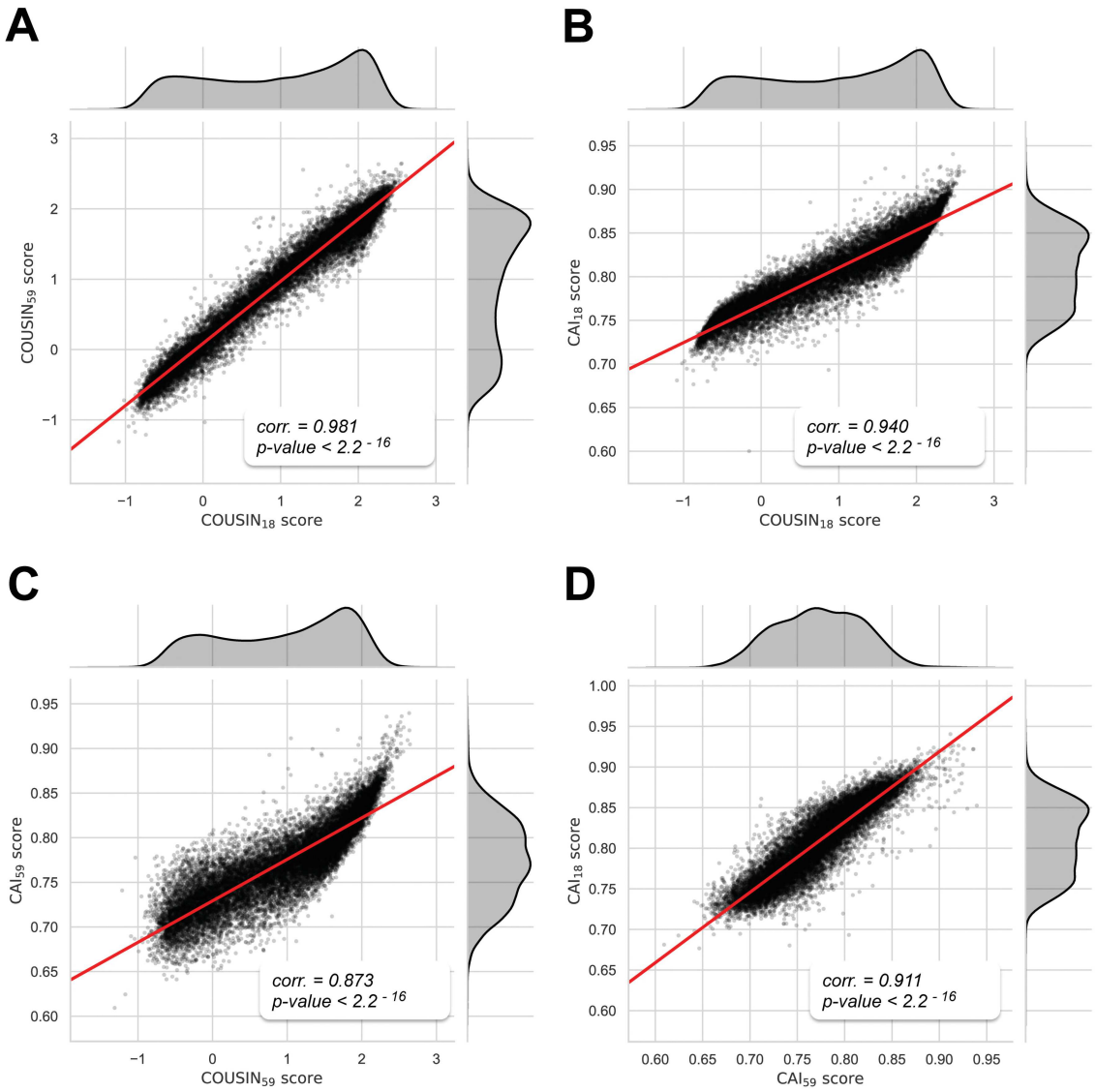

Figure 2.5 – **Analyse comparative des scores COUSIN et CAI obtenus pour les CDS d'*H. sapiens***. les sous-graphes A et D comparent les différentes versions des scores COUSIN (A) et CAI (D) alors que les sous-graphes B et C comparent les versions 18 (B) et 59 (C) des indices COUSIN et CAI. Les courbes flanquantes des nuages de points indiquent la densité observée des scores COUSIN ou CAI. À chaque comparaison, les résultats d'un test de corrélation de Pearson sont indiqués.

cette particularité génomique influence le CUB des gènes qu'on y retrouve [36, 43]. Par ailleurs, il existe chez la poule des macrochromosomes et microchromosomes (définis par une taille inférieure à 40Mb) qui possèdent un contenu global en GC divergeant : les microchromosomes sont enrichis en GC par rapport aux macrochromosomes (Figure 10b de l'article de *Hillier et al.* [229]).

Les résultats de la Figure 2.6 montrent la répartition des scores COUSIN en fonction de leur position dans les chromosomes 1 et 14 d'*H. sapiens*. Au sein des différents isochores, Les CDS forment des groupes ayant le même contenu en GC et le même score COUSIN, ce qui demeure en parfait accord avec les hypothèses décrivant un rôle majeur du biais mutationnel local sur le CUB [36, 43]. De tels résultats démontrent encore une fois l'efficacité de l'indice COUSIN à démarquer les différentes populations de CDS au sein d'un organisme, sans que l'indice CAI ne soit capable de déterminer avec précision de telles particularités. La bimodalité observée avec COUSIN chez *H. sapiens* s'explique donc par la position des gènes dans le génome : chaque mode de la distribution représente les CDS situés dans les isochores riches en AT ou en GC. Lorsque l'on observe la distribution du contenu en GC et des scores COUSIN selon le chromosome chez *H. sapiens* et *G. gallus*, on remarque une corrélation négative entre ces deux variables et la taille du chromosome (Figure 2.7). Cette corrélation est plus marquée chez la poule, chez qui une véritable scission est observée entre les micro et les macrochromosomes. De ce fait, la bimodalité observée chez *H. sapiens* et *G. gallus* est aussi associée aux différent CDS contenus dans les chromosomes (et dont le contenu en GC varie). De tels résultats sont à pondérer de par la forte variance observée au sein des chromosomes, mais permettent tout de même d'expliquer la bimodalité observée chez les deux organismes.

2.4 Conclusion

COUSIN est un indice novateur de mesure du CUB qui implique une détermination en fonction non seulement d'une référence, mais aussi d'une Hypothèse Nulle H_0 décrivant une équiprobabilité d'usage des codons. La gamme des résultats qu'il propose, sa fiabilité et sa simplicité en font un indice qui a toute sa place dans l'univers de la mesure du biais d'usage des codons. Grâce aux valeurs clefs de COUSIN (*i.e.* les valeurs 0 et 1), il est possible de considérer avec facilité la relation du CUB entre un gène et sa référence, mais aussi de comparer le CUB entre, à titre d'exemple, des gènes orthologues, car la significativité qualitative des scores obtenus ne varie pas en fonction de l'organisme considéré. Le score COUSIN d'une séquence peut être directement calculé grâce à l'outil `COUSIN`. Ce dernier s'illustre par sa simplicité d'utilisation et les fonctionnalités qu'il propose, allant de l'optimisation des gènes jusqu'à la création d'une CUT en passant par l'agrégation de données selon des approches de k-means et de clustering hiérarchique. L'analyse comparative effectuée au sein de cette étude témoigne des avantages de COUSIN sur son homologue CAI. Les scores COUSIN obtenus sont non-seulement plus affins avec le contenu en GC3 que ne le serait les scores CAI, mais semblent mieux déterminer les spécificités des organismes que ce dernier. Le choix du CUB moyen d'un organisme en guise de référence est ici purement démonstratif. Les objectifs ne sont pas de déterminer les tenants et aboutissants du CUB d'un organisme, mais bien de faire valoir les avantages et les désavantages

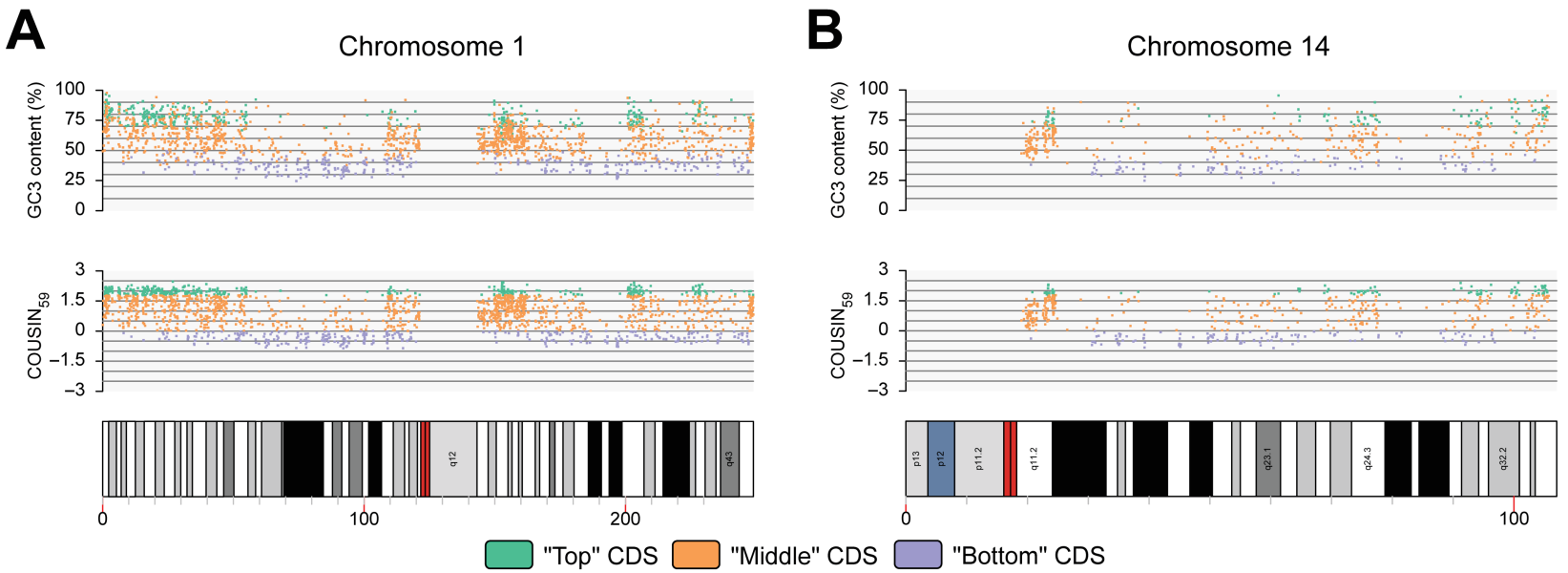

Figure 2.6 – **Contenu en GC3 (panneau supérieur), score COUSIN (panneau du milieu) et informations structurelles (panneau inférieur) des CDS des chromosomes 1 (A) et 14 (B) d'*H. sapiens*.** Les couleurs indiquent les CDS ayant un score COUSIN élevé (vert), moyen (orange) et bas (violet). L'abscisse représente la position des CDS le long du chromosome étudié (avec des valeurs associées représentant la distance du bras p au bras q en mégabases). Sur le panneau inférieur, le jeu de couleur indique les centromères (rouge), les isochores (ayant une gamme allant du blanc au noir pour représenter le contenu en GC des isochores : blanc richen en GC, noir riche en AT) et d'autres particularités telles que les constrictions secondaires (bleu).

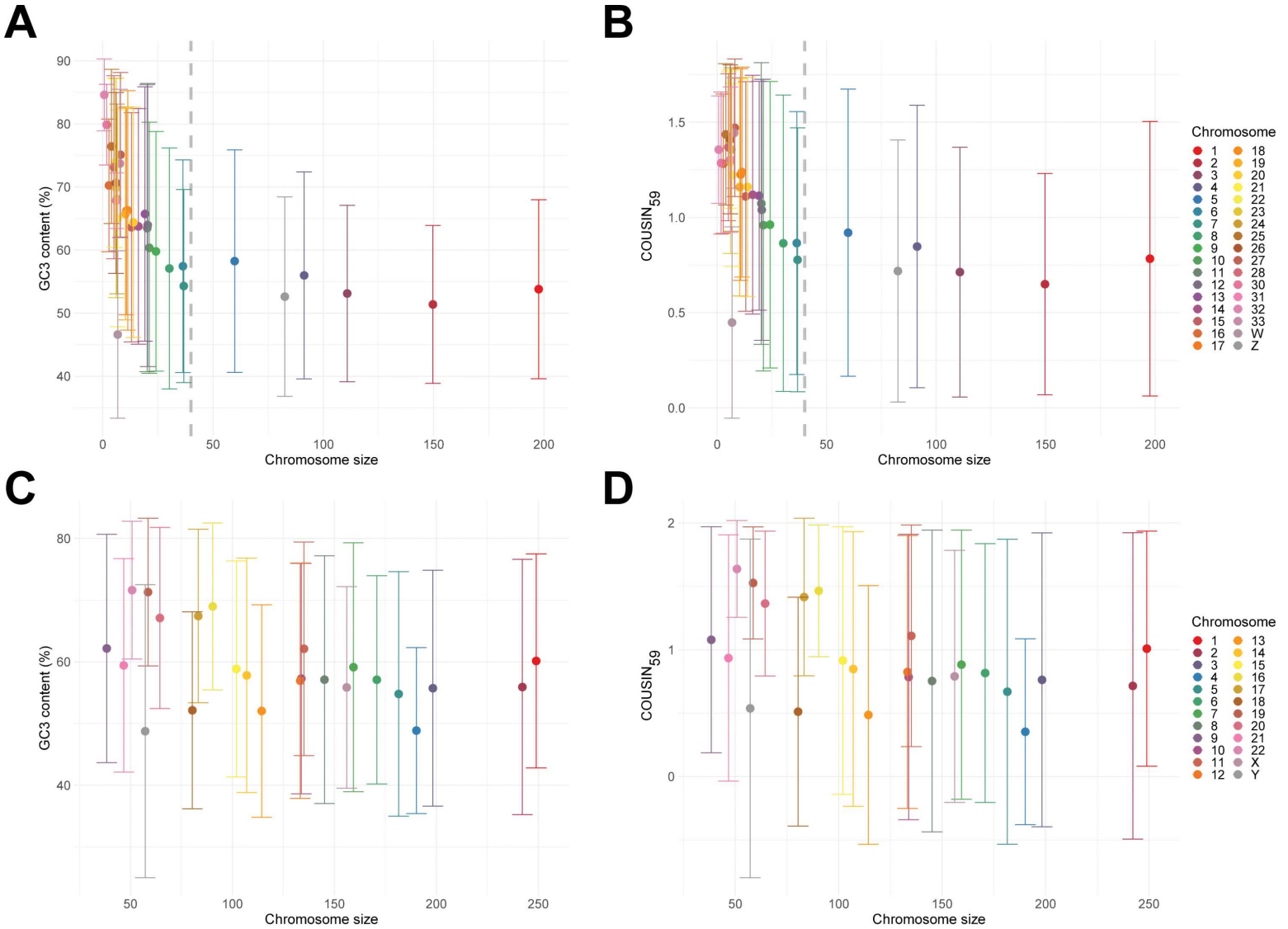

Figure 2.7 – **Valeurs Huber-M pour le contenu en GC3 (A et C) et la valeur COUSIN (B et D) en fonction de la taille des chromosomes chez *G. gallus* (A et B) et *H. sapiens* (C et D).** Chaque couleur représente un chromosome dont l'indicatif est donné en légende. Les lignes verticales données pour chaque chromosome représentent les valeurs MAD associées au score Huber-M [228].

de chaque indice face à une référence globale. Bien entendu, chaque organisme possède des particularités dans son CUB qui doivent être prises en compte dans toute analyse et ce notamment dans la construction d'un jeu de données de référence.

À l'avenir, COUSIN sera utilisé dans une série d'études visant à déterminer les spécificités dans le CUB de gènes viraux et d'organismes cellulaires. Au sein de cette thèse, nous utilisons COUSIN pour déterminer le CUB de gènes paralogues chez les Vertébrés puis de gènes appartenant aux polyomavirus humains.

Chapitre Trois

ÉVOLUTION DU CUB ET SOUS-FONCTIONNALISATION CHEZ LES GÈNES PARALOGUES : EXEMPLIFICATION PAR LES POLYPYRIMIDINE TRACT BINDING PROTEINS (*PTBP*)

3.1 Introduction

Les gènes homologues sont des séquences qui partagent une origine commune par événement de spéciation (orthologie) ou de duplication (paralogie) [230]. L'émergence de paralogues par duplication relâche les contraintes évolutives sur au moins une des deux copies du gène. Par processus évolutif, il est possible de voir apparaître de nouvelles fonctions chez l'un des deux paralogues, alors que l'autre conserve sa fonction initiale. Il est aussi possible que, malgré la conservation d'une fonction similaire chez les deux paralogues, l'un d'entre eux se spécialise et gagne des affinités avec, à titre d'exemple, un substrat particulier [231].

Le point de départ de notre recherche sont les observations expérimentales de *Robinson et al.* qui indiquent une expression différentielle des gènes paralogues de la « Polypyrimidine Tract Binding Protein » (*PTBP*) en fonction de leur composition nucléotidique [224]. Le génome des Vertébrés encode trois versions in-paralogues de ce gène qui possèdent toutes une fonction similaire : elles forment une classe de ribonucléprotéines appelées les « hnRNP RNA-Binding Proteins » et sont impliquées dans la modulation des événements d'épissages alternatifs chez les ARNm [223]. Au sein du même génome, les trois versions du gène *PTBP* affichent une forte similarité dans leur séquence protéique, aux alentours de 70 % chez *H. sapiens* [223].

Malgré leur forte ressemblance à l'échelle protéique, les trois gènes *PTBP* diffèrent grandement dans leur composition nucléotidique, leur CUB et leurs motifs d'expression. Chez l'Homme, *PTBP1* est enrichi en GC3 (% de GC à la troisième position du codon) et est exprimé dans tous

les tissus, alors que les gènes *PTBP2* et *PTBP3* sont riches en codons AT3 et affichent une forte expression dans le cerveau et dans les cellules hématopoïetiques (Figure Annexe C.1). *Robinson et al.* ont étudié l'expression de ces gènes dans des cellules humaines HeLa et HEK293, en plaçant tour à tour les gènes sous un même promoteur. Ils ont démontré que le paralogue *PTBP1* est plus fortement exprimé que ses deux paralogues AT3-riches, mais que l'expression du paralogue *PTBP2* pouvait être améliorée par la modification des codons synonymes vers un enrichissement en GC3 [224].

Au sein de cette étude, nous avons déterminé l'histoire évolutive des gènes *PTBP* et avons approfondi l'analyse du CUB aux Vertébrés. Pour cela, nous déterminons l'histoire évolutive des trois paralogues au sein de 74 espèces de Vertébrés et la comparons avec la diversité en CUB observée, en prenant soin de fractionner notre analyse aux individus mammifères et non-mammifères. Le CUB des trois paralogues est ensuite confronté à son contexte génomique au sein de quinze espèces, et ce dans l'objectif de déterminer la fraction expliquée par le biais mutationnel sur le CUB observé. Pour finir, nous mesurons la direction des mutations synonymes et non-synonymes pour y détecter des motifs expliquant la diversité en CUB observée au sein des trois paralogues. Nos résultats suggèrent que la diversité paralogue-spécifique du CUB chez les mammifères correspond à un processus de sous-fonctionnalisation par des motifs d'expression différentielle tissu-spécifique chez les *PTBP*.

3.2 Matériel et Méthodes

3.2.1 Construction du jeu de données de séquences

Nous avons construit un jeu de données de séquences nucléotidiques sur 47 mammifères, 27 Vertébrés non-mammifères et trois protostomiens en utilisant l'outil `BLAST` de la base de données NCBI [227]. Pour effectuer une telle recherche, nous nous sommes servis des paralogues *PTBP* humains comme jeu de données de référence (Tableau Annexe C.1 pour les numéros d'accès). Nous avons pu identifier les trois gènes chez tous les Vertébrés analysés, à l'exception du lapin de garenne *Oryctolagus cuniculus*, auquel il manquait *PTBP1* et du Xénipe grimpeur *Acanthisitta chloris*, auquel il manquait *PTBP3* (Tableau Annexe C.1). Le jeu de données final possédait 75 *PTBP1*, 76 *PTBP2* et 75 *PTBP3*. En tant qu' « outgroups » (éléments extérieurs phylogénétiquement éloignés) pour cette analyse, nous avons sélectionné les gènes orthologues de trois organismes protostomiens, pour lesquels seule une version du gène *PTBP* existe (Tableau Annexe C.1). À partir du jeu de données initial, nous avons identifié un sous-ensemble de neuf mammifères et de six non-mammifères pour lesquels nous avons effectué une analyse sur le contexte génomique des gènes. En effet, ces espèces possèdent des génomes relativement bien annotés, ce qui nous a permis d'extraire les informations sur les régions flanquantes et les introns des *PTBP* (Tableau Annexe C.2). À cause d'artefacts lors de l'annotation des gènes, ces régions non-codantes étaient manquantes chez certains *PTBP* de l'éléphant d'Afrique *Loxodonta africana*, du Gekko japonais *Gekko japonicus* et du requin baleine *Rhincodon typus*. Pour chacune des 15 espèces de l'analyse du contexte génomique, nous avons calculé les valeurs CAI [47] et COUSIN [94] grâce à l'outil en ligne `COUSIN` (accessible via le lien http://cousin.ird.fr).

3.2.2 Agrégation des *PTBP* selon leur CUB

Nous avons calculé la composition en codons ainsi que le CUB de chaque paralogue *PTBP* grâce à l'outil `COUSIN` [94]. Pour chacun d'entre eux, nous avons construit un vecteur de 59 positions contenant les fréquences relatives des codons synonymes. Nous avons par la suite effectué une Analyse en Composante Principale (ACP) sur les 229 vecteurs à 59 dimensions puis avons effectué des agrégations de données selon la méthode du k-means et du clustering hiérarchique.

3.2.3 Alignement et analyses phylogénétiques

Pour générer des alignements robustes sans introduire d'artefacts dus à la forte distance évolutive entre les paralogues, nous avons procédé étape par étape :

i Nous avons aligné séparément et à l'échelle des acides aminés chaque *PTBP* pour les mammifères et les non-mammifères.

ii Pour chaque paralogue, nous avons fusionné les alignements comprenant les séquences de mammifères et de non-mammifère, de manière à obtenir un seul alignement par *PTBP*.

iii Nous avons combiné les trois alignements en un seul.

iv Nous avons aligné les séquences des protostomiens à l'alignement global des Vertébrés

Chaque étape d'alignement a été effectuée à l'aide du programme `MAFFT` [232]. L'alignement obtenu a été rétro-traduit pour obtenir un alignement à l'échelle des codons. Pour finir, nous avons nettoyé l'alignement final en utilisant le programme `Gblocks` [233].

L'inférence phylogénétique a été réalisée à l'échelle des acides aminés et des nucléotides grâce au programme `RAxML` (v8.2.9) avec un bootstrap de 1000 [234]. Pour les données nucléotidiques, nous avons utilisé des partitions sur les codons et avons appliqué le modèle GTR + G4. Pour les données au format protéique, nous avons appliqué le modèle LG + G4. Nous avons retrouvé en parallèle l'histoire évolutive des 77 espèces en utilisant l'outil en ligne TimeTree [235]. Les distances entre les différents arbres phylogénétiques ont été mesurées par le biais de l'indice de Robinson-Foulds, qui se base sur la topologie [236], et l'approche du score de K-tree, qui prend en compte les différences à l'échelle de la topologie et de la longueur de branches [237]. Après inférence phylogénétique à partir des deux jeux de données, nous avons calculé l'état ancestral des ancêtres communs les plus proches de chaque clade de *PTBP* à l'aide de `RAxML`. À partir de ces états ancestraux, nous avons déterminé le nombre de mutations synonymes et non-synonymes de chaque individu à partir de l'état ancestral du clade qu'ils formaient.

3.2.4 Analyses statistiques

Les corrélations entre matrices ont été déterminées par le biais d'un test de Mantel. Des comparaisons non-paramétriques ont été produites en utilisant le test de Wilcoxon-Mann-Whitney (médiane des populations) et le test des rangs signés de Wilcoxon pour les comparaisons par paire. Les analyses statistiques ont été réalisés à l'aide des packages *ape* et *ade* de R et de JMP (v1.4.3.0).

3.3 Résultats

3.3.1 Les paralogues *PTBP* des Vertébrés diffèrent dans leur composition nucléotidique

Pour mieux comprendre l'histoire évolutive des *PTBP*, nous avons tout d'abord effectué une analyse du CUB et de la composition nucléotidique de ces gènes. D'un point de vue général, les gènes *PTBP1* sont enrichis en GC par rapport à *PTBP2* et *PTBP3* (moyennes respectives de 55,9 %, 42,3 % et 44,9 % pour le contenu en GC et 69,5 %, 33,4 % et 38,3 % pour le contenu en GC3 ; Figure 3.1, Tableau Annexe C.1). Par ailleurs, *PTBP1* possède une différence en GC3 entre les gènes de mammifères et de non-mammifères (respectivement 79,8 % contre 59,9 %). Un modèle de régression linéaire suivi par un test des étendues de Tukey sur les valeurs de GC3 avec comme variables explicatives les niveaux de paralogie (*i.e. PTBP1-3*), de taxonomie (*i.e.* mammifères et non-mammifères) ainsi que leur interaction identifient trois groupes principaux de *PTBP* (Tableau 3.1) : un premier rassemblant les *PTBP1* de mammifères, un deuxième les *PTBP1* de non-mammifères et un troisième comprenant tous les *PTBP2* et *PTBP3*. Le facteur le plus explicatif pour le contenu en GC3 est le niveau de paralogie *PTBP1-3*, comprenant pour lui seul 65% de la variance (Table 3.1). Ces tendances sont confirmées lors de la comparaison par paires entre les paralogues présents au sein du même génome mammifère, avec des différences significatives dans le GC3 selon l'ordre suivant : *PTBP1* > *PTBP3* > *PTBP2* (Test des rangs signés de Wilcoxon : *PTBP1* contre *PTBP2*, moyenne diff=48.0, S=539.50, p-value <0.0001 ; *PTBP1* contre *PTBP3*, moyenne diff=43.5, S=517.50, p-value <0.0001 ; *PTBP3* contre *PTBP2*, moyenne diff=4.5, S=406.50, p-value <0.0001). Il est important de noter que même si ces résultats sont tous significativement différents, les comparaisons par paires sur le GC3 entre *PTBP1* et *PTBP2-3* sont dix fois plus importantes que pour les comparaisons par paires entre *PTBP2* et *PTBP3*.

La distribution des résidus entre les valeurs observées et attendues de notre modèle nous permet d'identifier un nombre d' « outliers » (de valeurs aberrantes) avec des motifs taxonomiques intéressants (Tableau 3.2). Chez les non-mammifères, les trois paralogues de la truite arc-en-ciel *Oncorhynchus mykiss* possèdent un fort taux en GC3 (entre 67% et 76%), avec une différence significative par rapport aux valeurs prédites par le modèle (valeurs attendues entre 36% et 51%). Un cas similaire est observable chez le poisson-zèbre *Danio rerio* : les trois paralogues ont un contenu en GC3 avoisinant les 58%, ce qui est bien plus que les valeurs prédites par le modèle pour *PTBP2* et *PTBP3* (valeurs attendues aux alentours de 38%). De manière surprenante mais tout aussi intéressante, l'ornithorynque *Ornithorhynchus anatinus* et les trois marsupiaux de l'analyse, c'est à dire le diable de Tasmanie *Sarcophilus harrisii*, le koala *Phascolarctos cinereus* et l'opossum gris *Monodelphis domestica* ont une valeur de GC3 commune pour le paralogue *PTBP1*, aux alentours de 47%, ce qui est significativement plus bas que ce qui est prédit par le modèle (valeurs attendues d'environ 79%).

Au sein d'un certain nombre d'espèces Vertébrées, de fortes hétérogénéités dans la composition nucléotidique sont observées à l'échelle des chromosomes. Ces régions à forte dissimilarité sont appelées « isochores » [43, 107]. Pour explorer l'influence de l'environnement génétique sur la composition nucléotidique des *PTBP*, nous avons analysé le contexte génomique de 15 espèces

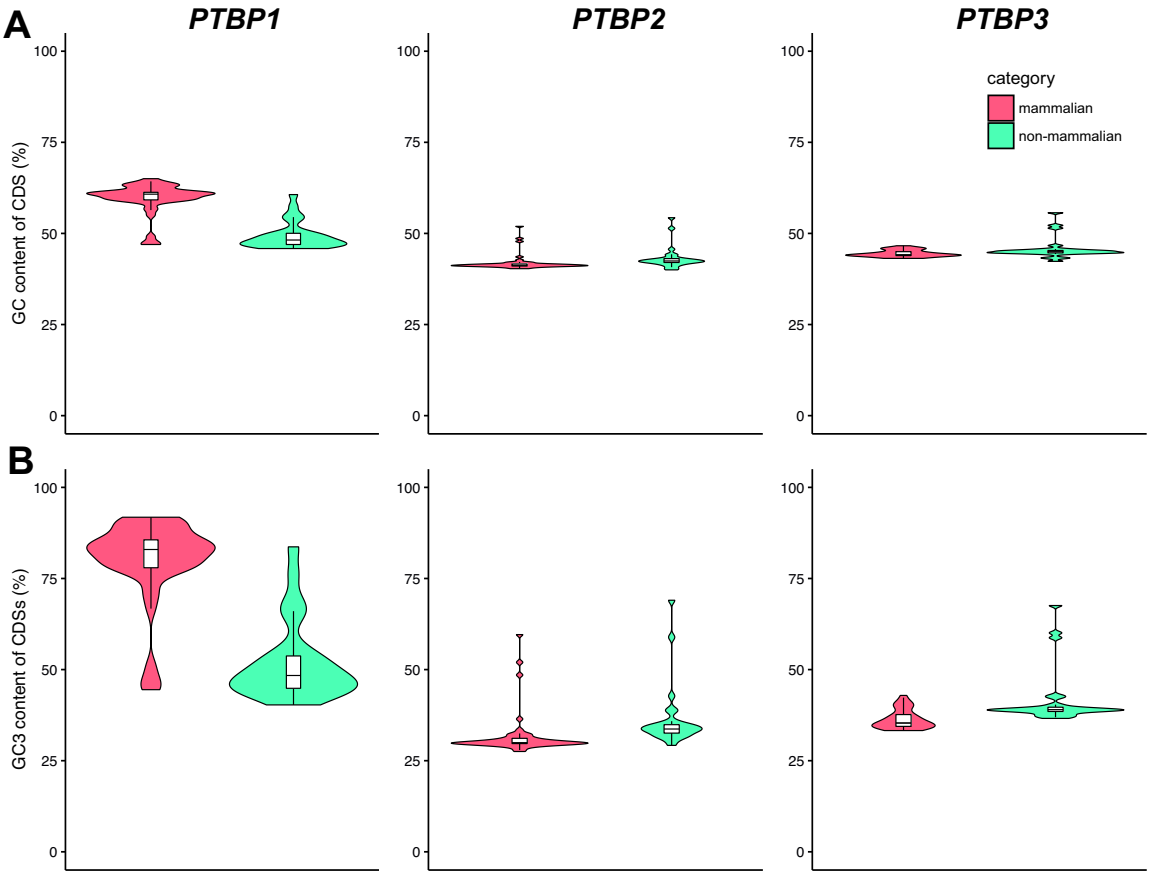

Figure 3.1 – **Contenu en GC (A) et en GC3 (B) des gènes *PTBP* chez les Vertébrés.** Les violin-plots affichent la distribution en GC et en GC3 tandis que les boxplot associés délimitent les valeurs de la médiane, des quartiles et des quantiles à 5% et 95% chez les différentes espèces de Vertébrés mammifères (rouge) et non-mammifères (bleu).

dont le génome est correctement annoté. Les variables utilisées au sein de cette analyse sont : i) le contenu en GC des régions introniques et flanquantes des *PTBP* ; ii) le GC3 de l'exome ; iii) le contenu en GC de tous les introns et de toutes les régions flanquantes au sein de chaque génome (Tableau 3.3 et Figure 3.2). Tout d'abord, pour *D. rerio*, la composition en GC3 des gènes *PTBP2* et *PTBP3* est clairement différente des autres espèces analysées et des tendances observées au sein de la Figure 3.2, et ce en concordance avec l'image d'« outlier » qu'il possède au sein du Tableau 3.2. Le poisson-zèbre a donc été exclu de la suite de l'analyse. Nous avons effectué une régression linéaire individuelle et étape par étape pour expliquer la variance en GC3 par la variation des valeurs de composition nucléotidique locales et globales décrites ci-dessus (Tableau 3.3). Pour chaque *PTBP*, le contenu en GC local explique le mieux les variations en GC3 des gènes, mais avec de fortes différences entre les paralogues : alors que la variation dans le contenu en GC des régions flanquantes et introniques corrèle fortement avec les variations de GC3 chez *PTBP1*

TABLE 3.1 – Modèle de régression linéaire et test des étendues de Tukey avec comme variable expliqué la composition en GC3 et en variable explicative les niveaux de paralogie (*PTBP1-3*), la taxonomie (*i.e.* mammifère ou non-mammifère) et leurs interactions. Les scores du modèle sont : Adj Rsquare=0.83 ; F ratio=205.7 ; Prob > F : <0.0001. Effets individuels des niveaux : i) paralogie : F ratio=274.3 ; Prob > F : <0.0001 ; ii) taxonomie : F ratio=27.2 ; Prob > F : <0.0001 ; iii) interaction paralogie*taxonomie : F ratio=87.9 ; Prob > F : <0.0001.

Niveau	**Least Sq. Mean (GC3%)**	**Écart-type**	**Groupe de Tukey**
Paralogie			
PTBP1	65.87	1.00	A
PTBP3	39.00	1.01	B
PTBP2	34.03	1.00	C
Taxonomie			
mammalian	49.32	0.70	A
non-mammalian	43.28	0.92	B
Paralogie*Taxonomie			
PTBP1, mammifère	79.81	1.22	A
PTBP1, non-mammifère	51.93	1.59	B
PTBP3, non-mammifère	41.64	1.62	C
PTBP3, mammifère	36.36	1.22	C, D
PTBP2, non-mammifère	36.27	1.59	C, D
PTBP2, mammifère	31.79	1.20	D

TABLE 3.2 – Valeurs « outliers » des gènes en accord avec les valeurs attendues du modèle de régression linéaire sur les niveaux de paralogie (*PTBP1-3*), de taxonomie (mammifère ou non-mammifère) et de leur interaction.

Espèce	**Paralogue**	**GC3 observé (%)**	**GC3 attendu (%)**	**déviation du GC3 (%)**
mammifère				
Desmodus rotundus	*PTBP2*	59.60	31.79	27.81
Miniopterus natalensis	*PTBP2*	48.52	31.79	16.72
Monodelphis domestica	*PTBP1*	44.49	79.81	-35.32
Ornithorhynchus anatinus	*PTBP1*	51.14	79.81	-28.67
Ornithorhynchus anatinus	*PTBP2*	52.00	31.79	20.21
Phascolarctos cinereus	*PTBP1*	47.53	79.81	-32.28
Sarcophilus harrisii	*PTBP1*	45.44	79.81	-34.37
non-mammifère				
Danio rerio	*PTBP2*	58.89	36.27	22.62
Danio rerio	*PTBP3*	60.08	41.64	18.44
Lepisosteus oculatus	*PTBP3*	58.73	41.64	17.10
Oncorhynchus mykiss	*PTBP1*	76.27	51.93	24.34
Oncorhynchus mykiss	*PTBP2*	69.03	36.27	32.76
Oncorhynchus mykiss	*PTBP3*	67.58	41.64	25.95
Pogona vitticeps	*PTBP1*	83.68	51.93	31.75

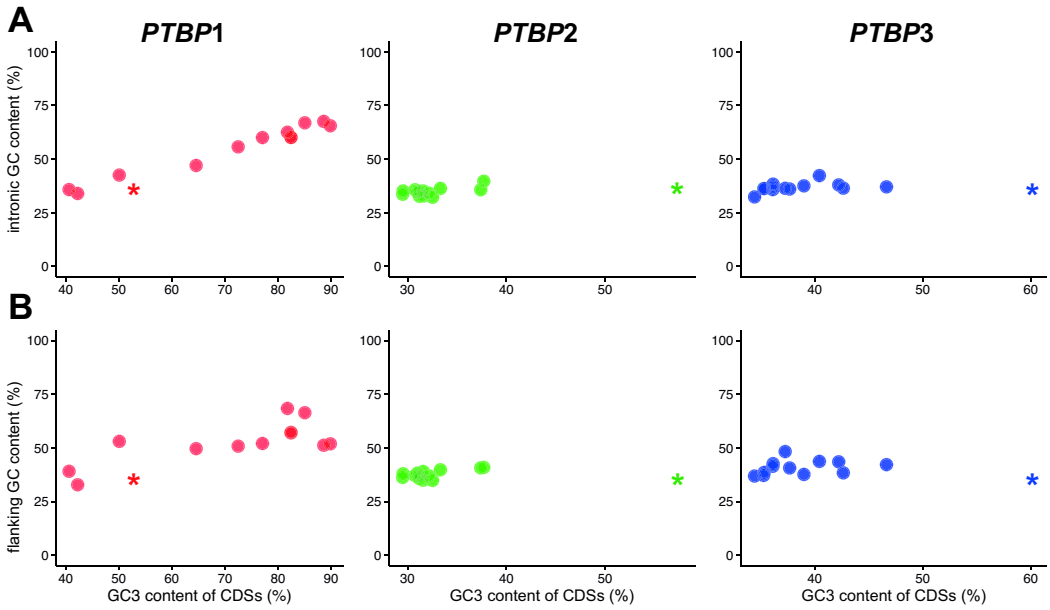

Figure 3.2 – **Variation dans le contenu en GC3 (abscisse) et dans le contenu intronique (A, ordonnée) et des régions flanquantes (B, ordonnée) chez les *PTBP***. Chaque point représente un des 15 espèces de l'analyse sur le contexte génomique. L'astérisque indique les valeurs obtenues pour l'espèce *D. rerio*, qui montre des résultats particuliers pour *PTBP2* et *PTBP3*, en accord avec son comportement « outlier » dans les modèles.

(R^2=0.97), et d'une manière moins prononcée chez *PTBP3* (R^2=0.78), la fraction de variance expliquée par ces variables chute de manière significative chez *PTBP2* (R^2=0.46).

3.3.2 Les paralogues *PTBP* diffèrent dans leur CUB

Pour chaque CDS *PTBP*, nous avons extrait les fréquences relatives des codons synonymes et avons appliqué différentes approches pour réduire l'information et visualiser les tendances en CUB. Les résultats d'une Analyse en Composante Principale (ACP) sont montrés au sein de la Figure 3.3. Le premier axe capture 68,9 % de la variance, bien plus que les deuxièmes et troisièmes axes (respectivement 6,7 % et 3,2 %). Au sein des familles de codons de multiplicité 2, les deux codons sont forcément symétriques, ce qui crée une redondance dans la détermination du CUB. Pour corriger cela, nous avons simplifié l'ACP en n'analysant que les familles de codons de multiplicité 4 et 6 (Figure Annexe C.2). Les résultats obtenus demeurent néanmoins similaires à l'ACP décrite au sein de ce rapport. Ici, le premier axe rassemble les codons par leur composition en GC3, avec l'exception du codon UUG-Leu, qui se groupe avec les codons riches en AT3. Ce premier axe différencie clairement les gènes *PTBP1* des mammifères de tous les autres gènes *PTBP2* et *PTBP3*. Les *PTBP1s* appartenant aux non-mammifères se répartissent entre les *PTBP1* de mammifères et les *PTBP3*, mais se rapprochent aussi des gènes *PTBP* des protostomiens. Le

deuxième axe de l'ACP décrit des motif abstraits dans le rassemblement des codons : i) une séparation entre les codons se terminant par G ou C, mais pas dans les codons se terminant par A ou T ; ii) Une forte contribution antagoniste entre les codons AGA et AGG de l'arginine. Ce second axe sépare toutefois les *PTBP2* des *PTBP3*, en accord avec les tendances en composition nucléotidiques décrites précédemment : une comparaison par paires indique que les *PTBP3* sont enrichis en codons se terminant par C par rapport à *PTBP2*, avec des valeurs de 21, 7 % contre 15,4 % (test des rangs signés de Wilcoxon : moyenne diff=6.2, S=1184.0, p-value <0.0001).

Pour approfondir notre recherche de particularités au sein des *PTBP*, nous avons effectué une agrégation de données hiérarchique ainsi qu'une agrégation selon la méthode des k-means.

TABLE 3.3 – Résultats d'une analyse par la méthode des moindres carrées (approche individuelle ou séquentielle) pour expliquer les variations en GC3 des gènes *PTBP* par les variations globales et locales dans la composition en nucléotides au sein de 14 génomes correctement annotés. Pour chaque gène, les variables individuelles sont ordonnées en fonction de leur contribution au modèle séquentiel. Les variables avec la mention « N.S » (Non-significatif) ne contribuent pas avec une puissance explicative quand ajoutées au modèle séquentiel.

PTBP1				
	Contribution individuelle		**Contribution séquentielle**	
Paramètre	**R^2**	**BIC**	**R^2**	**BIC**
Local intronic GC	0.96	74.42	0.96	74.42
Global intronic GC	0.03	111.98	0.97	71.23
Global flanking GC	0.05	111.70	0.98 (N.S.)	72.26
Global exomic GC3	0.62	100.71	0.98 (N.S.)	74.27
Local flanking GC	0.55	112.66	0.98 (N.S.)	76.55
PTBP2				
	Contribution individuelle		**SContribution séquentielle**	
Paramètre	**R^2**	**BIC**	**R^2**	**BIC**
Local flanking GC	0.46	60.12	0.46	60.12
Global flanking GC	0.03	67.66	0.49 (N.S.)	61.86
Local intronic GC	0.37	61.95	0.49 (N.S.)	64.38
Global exomic GC3	0.09	66.75	0.49 (N.S.)	66.89
Global intronic GC	0.05	67.38	0.50 (N.S.)	69.35
PTBP3				
	Contribution individuelle		**Contribution séquentielle**	
Paramètre	**R^2**	**BIC**	**R^2**	**BIC**
Local intronic GC	0.78	78.11	0.78	78.11
Global intronic GC	0.12	96.38	0.80 (N.S.)	79.56
Global exomic GC3	0.02	97.73	0.82 (N.S.)	80.66
Local flanking GC	0.38	91.77	0.84 (N.S.)	81.70
Global flanking GC	0.02	97.77	0.84 (N.S.)	84.27

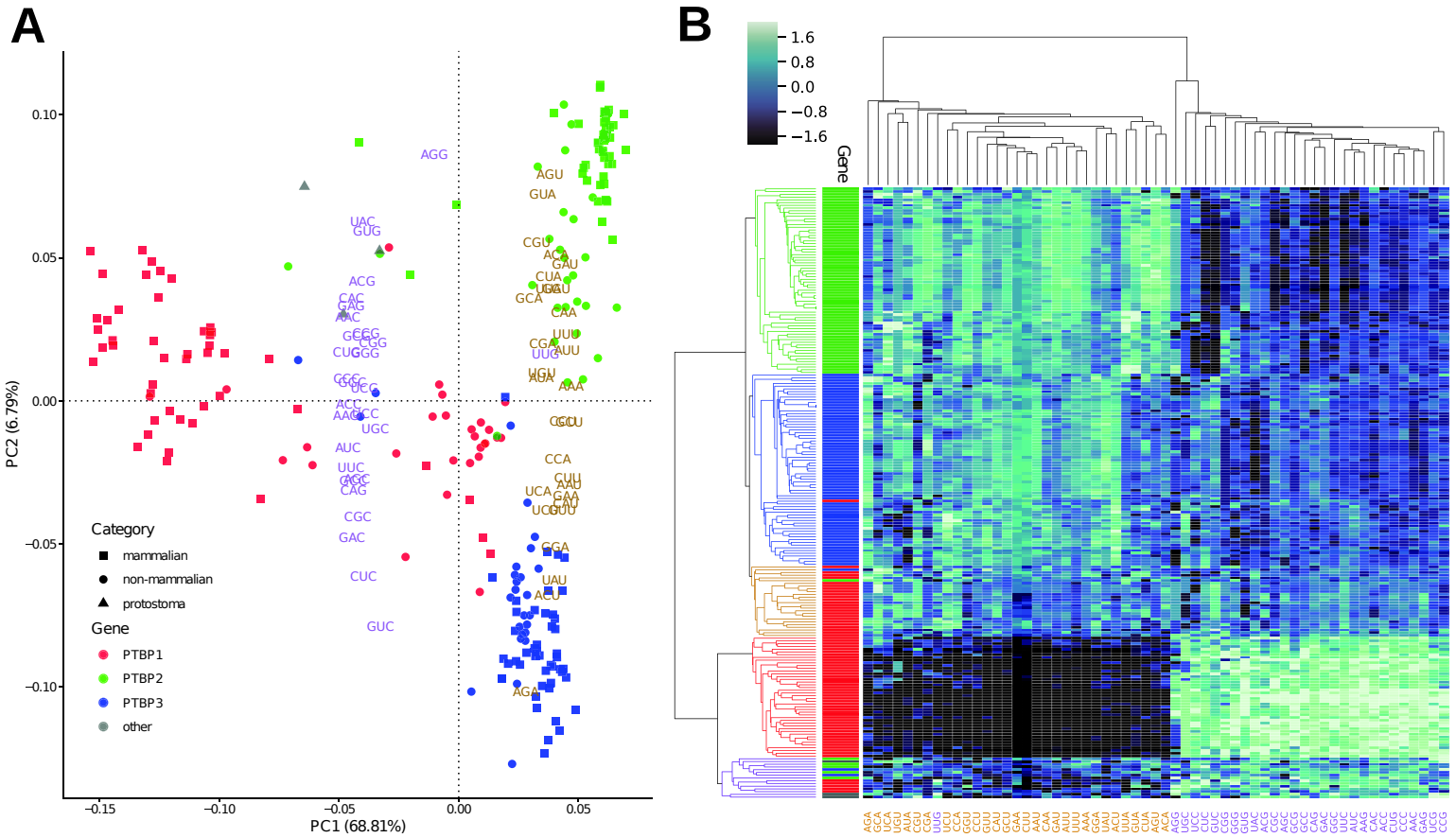

Figure 3.3 – **Analyse du CUB des *PTBP***. A) Deux premières dimensions d'une ACP basée sur le CUB des *PTBP1* (en rouge) *PTBP2* (en vert),*PTBP3* et des gènes protostomiens (en gris). L'information taxonomique est donnée par un jeu de formes : mammifères (carrés), non-mammifères (ronds) et protostomiens (triangle). L'ACP a été réalisée par l'analyse des vecteurs de 59 dimensions représentant les fréquences relatives des codons synonymes de chaque gène. Les valeurs propres des matrices sont données par leur position dans le graphe. Chaque variable est identifiée par son nom et par un code couleur pour les codons se terminant en GC (violet) et en AT (orange). Le % de variance expliquée pour chaque axe est donné entre parenthèses. B) Heatmap sur les individus de cette analyse (ligne) et sur le vecteur de 59 positions qui leur est associé (colonne). Les dendrogrammes ont été construits par le biais d'un clustering hiérarchique, et les couleurs de l'arbre des individus (à gauche) correspondent aux groupes formés à partir du clustering. La barre de côté indique la nature du gène correspondant à la ligne : *PTBP1* (en rouge), *PTBP2* (en vert) et *PTBP3* (en bleu).

Chaque analyse rassemble les *PTBP* par le contenu en GC3 qu'ils possèdent. Le dendrogramme des *PTBP* issu du clustering hiérarchique (lignes de la figure 3.3 B) indique l'existence de cinq clades qui regroupent les *PTBP* suivants : *PTBP1* de mammifères, *PTBP1* de non-mammifères, *PTBP2*, *PTBP3* et un cinquième groupe contenant les *PTBP* des protostomes ainsi que quelques individus de chaque paralogue (score de consistence de Kappa-Fleiss = 0.76). Cette même méthode de clustering stratifie les codons selon leur contenu en GC3 : d'un côté les codons enrichis en GC3 et de l'autre ceux en AT3, avec la claire exception du codon UUG de la leucine, qui se groupe au milieu des codons enrichis en AT3 (voir le dendrogramme en colonne de la Figure 3.3 B). L'approche de détection du nombre optimal de groupes du k-means (*i.e.*, la méthode du coude) identifie quatre groupes qui séparent les paralogues de la manière suivante : *PTBP1*, *PTBP2*, *PTBP3* et un dernier groupe contenant les protostomes ainsi que des individus appartenant à chaque paralogue (score de consistence de Kappa-Fleiss = 0.75).

Globalement, les méthodes d'agrégation des données hiérarchique et en k-means, toutes les deux basées sur les vecteurs à 59 dimensions des fréquences des codons synonymes, sont en accord avec l'autre (score de consistence de Kappa-Fleiss = 0.83), et sont en concordance avec l'ACP produite. Le CUB définit donc des groupes de gènes consistants avec leur orthologie et leur taxonomie. Il est intéressant de noter que pour certaines espèces, les trois paralogues possèdent une distribution unique du CUB, comme un CUB commun chez les trois paralogues du requin baleine *R. typus* ou encore des compositions inattendues dans le CUB de la chauve-souris *Miniopterus natalensis*.

Nous avons analysé le CUB des gènes *PTBP* des 15 espèces pour lesquelles nous avions une annotation correcte du génome, en utilisant comme référence le CUB moyen des génomes associés (Tableau 3.4). Nos résultats soulignent de fortes différences chez les paralogues de mammifères : *PTBP1* possède des valeurs COUSIN au dessus de 1 alors que *PTBP2* et *PTBP3* des valeurs en dessous de 0. En suivant la signification des scores COUSIN [94], ces résultats démontrent que les *PTBP1* des mammifères sont enrichis en codons retrouvés en forte proportion au sein du génome, alors que les *PTBP2-3* sont enrichis en codons rares.

3.3.3 Reconstruction phylogénétique des *PTBP*

Nous avons exploré la relation évolutive entre les *PTBP* par le biais d'une inférence phylogénétique à l'échelle des acides aminés et des nucléotides (Figure 3.4). Le jeu de données final contenait 74 *PTBP*s de mammifères (47 espèces pour 39 familles) et de Vertébrés non-mammifères (27 espèces pour 24 familles). Nous avons utilisé les gènes *PTBP* des trois protostomiens comme outgroups dans cette analyse. Chacune des deux phylogénies affiche trois grands clades qui représentent chacun un des paralogues. Au sein des deux topologies obtenues, les orthologues *PTBP1* et *PTBP3* se regroupent, mais il est à noter que la taille de la branche reliant les protostomiens aux Vertébrés est si démesurée qu'elle ne permet pas une identification correcte de la résolution inter-*PTBP* chez les Vertébrés. De manière générale, les deux phylogénies sont congruentes (Tableau 3.5). La forte distance entre les arbres nucléotidiques et protéiques pour *PTBP2* provient de désagréments chez les branches courtes, comme démontré par le score K-tree de cet orthologue (pour rappel, l'index de Robinson-Foulds ne permet qu'une analyse sur la topologie, alors que le score K-tree se base sur la topologie et la taille des branches). Dans tous

TABLE 3.4 – Régression linéaire et test des étendues de Tukey, avec comme variable expliquée les valeurs COUSIN de chaque *PTBP* (calculée avec comme référence le CUB moyen de l'organisme). Les variables explicatives sont les niveaux de paralogie (*PTBP1-3*), de taxonomie (*i.e.* (non-)mammifère) et leur interaction. Les scores du modèles sont de : Adj Rsquare=0.82 ; F ratio=36.84 ; Prob > F : <0.0001. Effets individuels des niveaux : i) paralogie : F ratio=40.72 ; Prob > F : <0.0001 ; ii) taxonomie : F ratio=10.87 ; Prob > F : =0.0021 ; iii) interaction paralogie*taxonomie : F ratio=28.11 ; Prob > F : <0.0001.

Level	**Least Sq. Mean (COUSIN)**	**Std. err.**	**Tukey's HSD group**
Paralog			
PTBP1	1.45	0.11	A
PTBP3	0.29	0.11	B
PTBP2	0.19	0.11	B
Taxonomy			
mammalian	0.44	0.080	A
non-mammalian	0.85	0.098	B
Paralog*Taxonomy			
PTBP1, mammalian	1.90	0.14	A
PTBP1, non-mammalian	0.99	0.17	B
PTBP2, non-mammalian	0.81	0.17	B
PTBP3, non-mammalian	0.75	0.17	B
PTBP3, mammalian	-0.16	0.14	C
PTBP2, mammalian	-0.43	0.14	C

les cas, la structure interne des sous-arbres *PTBP1-3* récapitule correctement la taxonomie des espèces (Tableau 3.5). Certaines espèces identifiées par le modèle mathématique comme ayant une composition nucléotidique déviante possèdent des branches longues dans la reconstruction phylogénétique (par exemple le *PTBP3* de *O. mykiss*).

Nous avons ensuite analysé la correspondance entre les distances par paires des phylogénies nucléotidiques et protéiques. Nous observons une bonne entente entre les deux reconstructions, à l'exception des *PTBP2* des mammifères qui possèdent une faible divergence sur l'arbre protéique (Figure 3.5 B, Tableau Annexe C.5 B). Sur les *PTBP1*, le graphe montre un clair regroupement des mammifères monotrèmes et marsupiaux, qui se dégagent du reste des mammifères (et ce sur les deux phylogénies). Cette distribution est en accord avec le fait que les monotrèmes et les marsupiaux se détachent des mammifères placentaires au sein de la phylogénie des *PTBP1* (Figure 3.4). La même conclusion peut être tirée de l'ornithorynque, dont le *PTBP3* se détache du reste des mammifères. Pour les paralogues des mammifères, nous observons une augmentation du nombre de mutations, et plus particulièrement de mutations non synonymes chez *PTBP3* par rapport à *PTBP1*. Les motifs de mutations sont analysés en détail ci-dessous. Les histogrammes dénombrant une accumulation des mutations synonymes et non-synonymes confirment que les *PTBP1* de mammifères ont accumulé le plus grand nombre de mutations synonymes, en comparaison avec les *PTBP1* non-mammifères et les autres gènes (Figure Annexe C.4).

Pour finir, nous avons analysé la connexion entre les distances évolutives à l'échelle nucléotidique des paralogues *PTBP* et leur distances basées sur le CUB (Figure 3.5 A, Tableau Annexe C.5 A). Une tendance démontrant une augmentation des différences en CUB en fonction des distances évolutives n'est visible que pour les *PTBP1* et *PTBP3* des mammifères. Pour le *PTBP1* des mammifères, le graphe associé à cette étude différencie clairement deux populations à l'échelle évolutive et du CUB : l'une contenant les monotrèmes et les marsupiaux, l'autre les placentaires. Ce même graphe sépare clairement le *PTBP2* de l'ornythorinque et des chauves-souris *M. natalensis* et *Desmodus rotundus* des autres mammifères, mais aussi le *PTBP2* de la truite arc-en-ciel des autres *PTBP2* de Vertébrés non-mammifères. Pour finir, le *PTBP3* de l'ornythorinque semble lui aussi posséder un CUB et un profil nucléotidique divergeant par rapport aux autres mammifères. Chacune de ces déviations est en parfait accord avec les profils particuliers décrits ci-dessus par les modèles mathématiques de la composition nucléotidique de ces orthologues, à l'exception du *PTBP3* de l'ornithorynque.

3.3.4 Les *PTBP1* des mammifères accumulent des substitutions synonymes GC-enrichissantes

Nous avons observé que les gènes *PTBP1* sont généralement plus enrichis en GC (et plus particulièrement en GC3) que les autres paralogues au sein d'un même génome, et que cet enrichissement est bien plus présent chez les placentaires. Nous avons donc voulu évaluer si un motif mutationnel directionnel particulier souligne cet enrichissement, et ce au regard des mutations synonymes. Pour cela, nous avons inféré l'état ancestral des séquences pour chaque paralogue, avons déterminé les mutations synonymes et non-synonymes entre cet état ancestral et chaque individu apparenté et avons construit les matrices de mutation correspondantes (Tableau Annexe C.6). Les deux premiers axes d'une Analyse en Composante Principale sur ces matrices

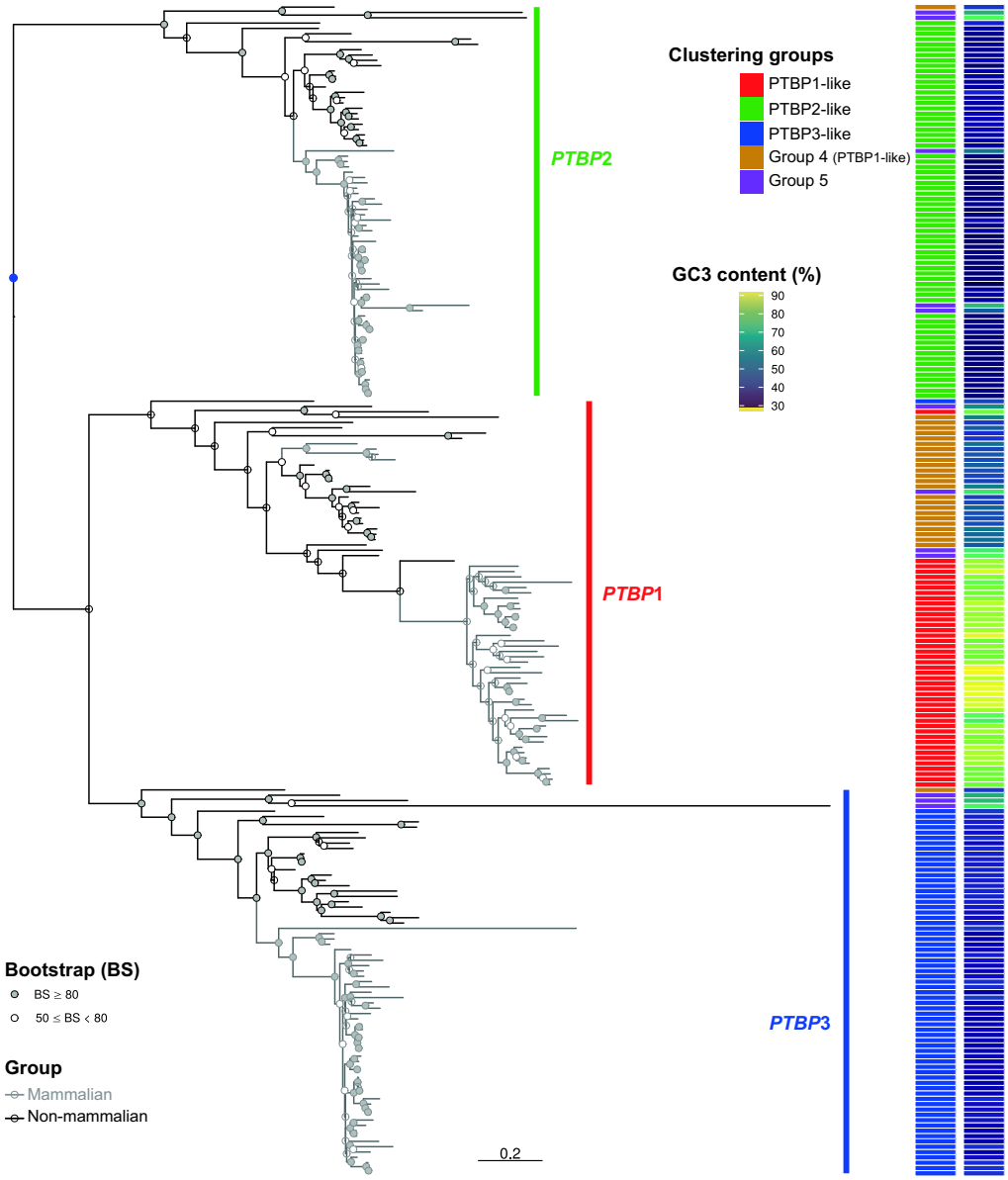

Figure 3.4 – **Phylogénie en maximum de vraisemblance des gènes *PTBP***. Le phylogramme représente les clades *PTBP2* (barre latérale verte), *PTBP1* (barre latérale rouge) et *PTBP3* (barre latérale bleue). Les gènes des protostomiens ne sont pas montrés au sein de cette figure, dans l'objectif de focaliser la représentation graphique sur les Vertébrés, mais leur placement sur l'arbre et la polarité qu'ils induisent est donnée par le point bleu. Les branches grises indiquent les résolutions à l'échelle des mammifères, et les noires à l'échelle des non-mammifères. Une attention particulière est notée sur les *PTBP1* de mammifères, où la résolution n'est pas monophylétique. Les points pleins indiquent une valeur de bootstrap supérieure à 80, alors que les points vides un bootstrap moins élevé. Les barres latérales de gauche indiquent la classification proposée par le clustering hiérarchique, avec le même code couleur que donné dans la Figure 3.3 B. La barre latérale de droite affiche le contenu en GC3 des gènes, avec un gradient allant de 0% (en bleu) à 100% (en jaune).

TABLE 3.5 – Comparaison entre l'arbre des espèces et les sous-arbres de la phylogénie obtenue à partir des données nucléotidiques. Chaque sous-arbre correspond à un paralogue *PTBP*. Le score K-tree les distances topologiques et par paire entre les arbres après avoir calibré les deux arbres sur leur taille. Le test de Robinson-Foulds compare la topologie entre les deux arbres. Les valeurs montrées correspondent à la fraction de nœuds divergeants entre les deux arbres.

Arbre de référence	**Arbre comparé**	**score K-tree**	**score de Robinson-Foulds**
Arbre nucléotidique *v.s.* arbre des espèces			
PTBP1	arbre des espèces	0.759	42
PTBP2	arbre des espèces	0.762	24
PTBP3	arbre des espèces	1.700	28
Arbre nucléotidique *v.s.* arbre protéique			
PTBP1-AA	*PTBP1*-NT	0.149	78
PTBP2-AA	*PTBP2*-NT	0.129	110
PTBP3-AA	*PTBP3*-NT	0.380	40

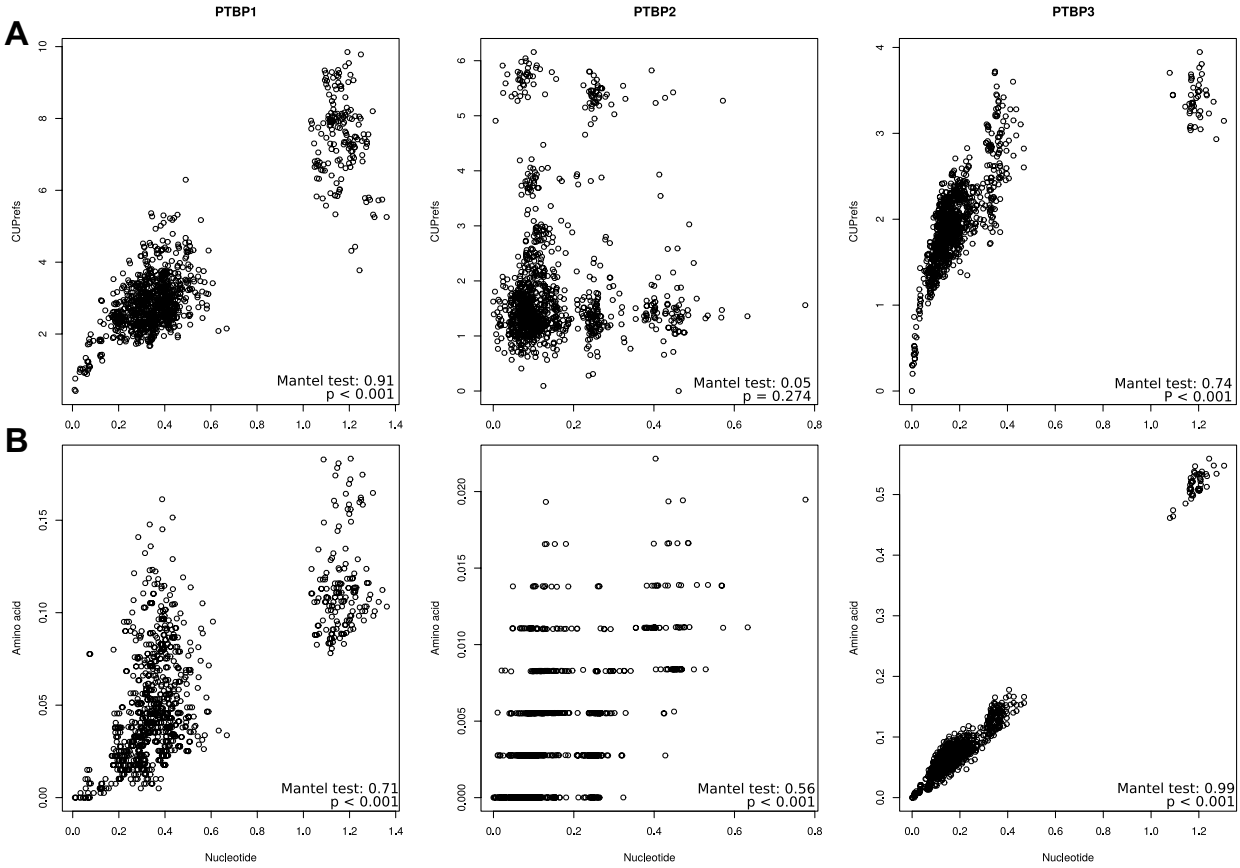

Figure 3.5 – **Distance par paire de l'arbre nucléotidique contre A) distances par paires en CUB et B) distances par paire de l'arbre protéique pour les *PTBP* de mammifères.** Les résultats d'un test de Mantel (sur la corrélation entre les variables) sont donnés pour chaque graphe.

mutationnelles capturent, à eux deux, 66,95 % de la variance entre individus (Figure 3.6). Le premier axe de cette ACP sépare les substitutions synonymes des non-synonymes. De manière intéressant, les transitions T<->C sont associées aux profil des mutations synonymes de l'ACP (comme attendu), alors que les transitions G<->A sont retrouvées au sein du profil des mutations non-synonymes. Le deuxième axe sépare les substitutions par leur effet sur la composition nucléotidique : stabilisation en GC dans un sens, stabilisation en AT dans l'autre. De manière surprenante, le spectre mutationnel des *PTBP1* de mammifères diffère largement du reste des paralogues. Les substitutions au sein de ces gènes, GC-enrichissantes pour les mutations synonymes et non-synonymes, définissent grandement ce deuxième axe. Au contraire, les mutations synonymes des gènes *PTBP3* et les mutations synonymes et non-synonymes des *PTBP2* ont tendance à être AT-enrichissantes. Pour finir, les tendances mutationnelles des mammifères sont radicalement différentes des non-mammifères pour *PTBP1*, alors que ces tendances ne sont pas vérifiées pour *PTBP2* et *PTBP3*, et ce quelle que soit la nature des substitutions.

3.4 Discussion

L'utilisation déséquilibrée des codons synonymes a, depuis sa découverte, soulevé bon nombre de questions au sein de la communauté scientifique. Leur analyse a permis l'élaboration de controverses fructifiantes entre les défenseurs de *tout est neutraliste* et de *tout est sélection*, et a ouvert la porte d'une recherche pour la détection de signaux et de codes au sein des motifs du CUB. Les études du CUB sont à deux niveaux. D'un côté, c'est sur son origine que l'on se questionne : à quel degré le CUB est-il le résultat d'un partage entre le biais mutationnel et les processus de sélection ? D'un autre côté, c'est plutôt sur son rôle que l'on se focalise : comment et à quel point un CUB particulier peut-il être lié à des processus de régulation de l'expression des gènes, que ce soit en modifiant les cinétiques et les dynamiques de la transcription de l'ADN, de la maturation et de la stabilité de l'ARNm, de sa traduction, ou encore du repliement des protéines et de leur stabilité ? Au cours de cette étude nous nous sommes basés sur les résultats expérimentaux proposés par *Robinson et al.* à propos de l'expression différentielle des *PTBP* humains en fonction de leur CUB [224]. À partir de cet exemple, nous avons exploré par raisonnement inductif la nature de la connexion entre l'évolution des gènes paralogues et celle de leur CUB. Nos résultats montrent que les trois paralogues *PTBP* des Vertébrés, qui possèdent des différences dans leur expression chez *H. sapiens*, ont aussi une composition nucléotidique et un CUB différent à l'échelle des Vertébrés. Nous proposons ici que ce motif évolutif est compatible avec un phénomène d'évolution phénotypique par sous-fonctionnalisation (dans ce cas, une spécialisation décrivant des différences d'expression selon le tissu), associé à une évolution à l'échelle génotypique par association à des motifs spécifiques du CUB.

Nous avons reconstruit les relations phylogénétiques et avons analysé l'évolution et la diversité du CUB chez les différentes versions des paralogues *PTBP* de 74 espèces de Vertébrés. La résolution phylogénétique indique que les événements de duplications semblent anciens : tous les Vertébrés semblent posséder les trois versions du gène, et celles-ci forment des clades distincts et phylogénétiquement éloignés. Cette observation est cohérente avec les informations des bases de données des orthologues et des paralogues ENSEMBL et ORTHOMAM [223, 238, 239].

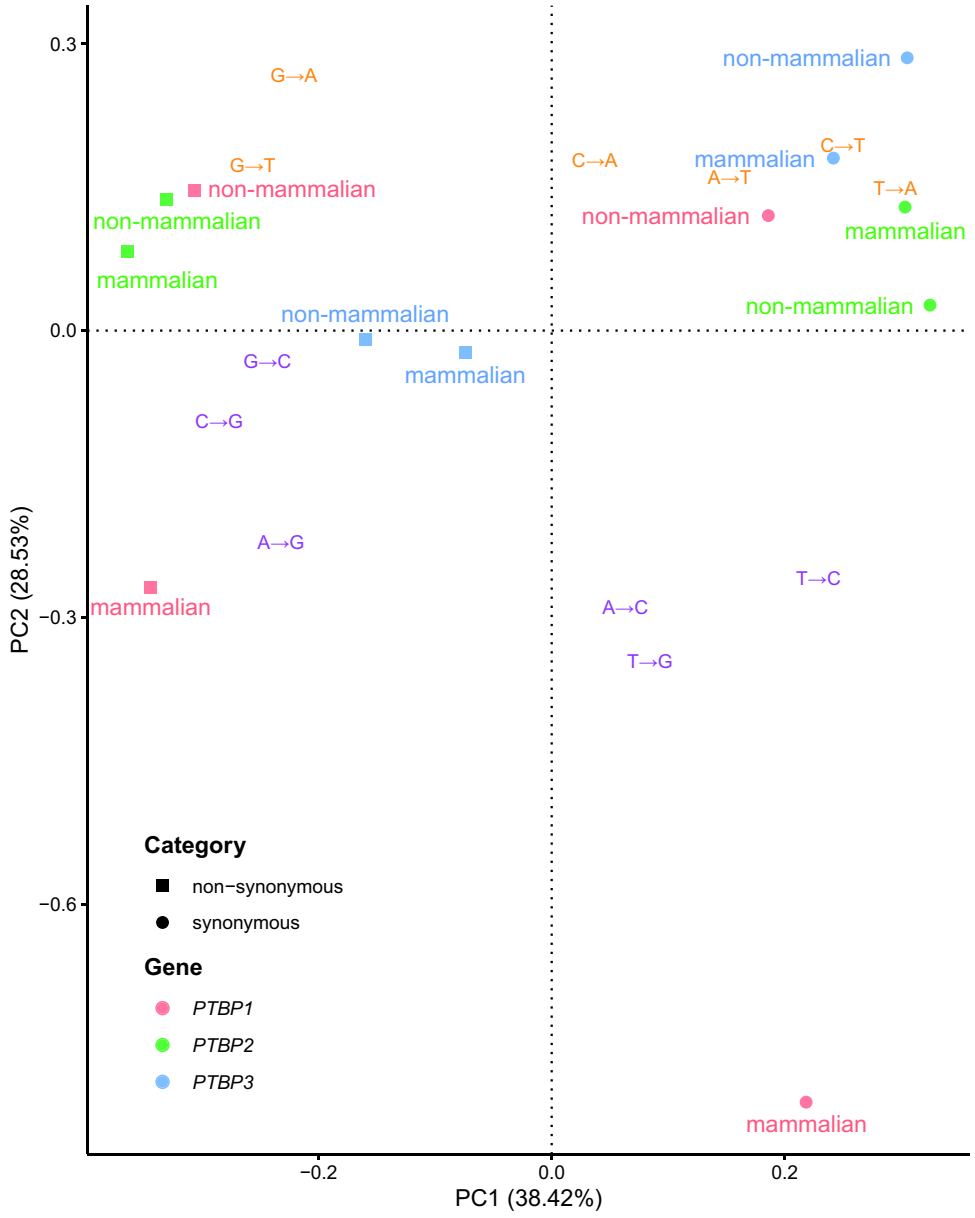

Figure 3.6 – **Spectre des substitutions synonymes et non-synonymes des *PTBP*s.** Cette ACP a été construite en utilisant les matrices de substitutions synonymes et non-synonymes pour chaque *PTBP*, elles-mêmes construites à partir de l'inférence phylogénétique et de la comparaison entre les individus et les états ancestraux de celle-ci. Les variables de cette ACP sont les natures des substitutions (*e.g.* A->G), identifiées par un code couleur représentant les mutations GC-enrichissantes / stabilisantes (violet) our AT-enrichissantes / stabilisantes (orange). La position des variables est représentée en fonction de leur valeurs propres. Les individus de cette ACP sont les catégories de mutations des *PTBP*, stratifiées par leur nature (synonyme ou non-synonyme), par leur orthologie (code couleur donné dans la figure pour les différents *PTBP*). et leur taxonomie (mammifère ou non-mammifère).

Bien que nos résultats suggèrent que *PTBP1* et *PTBP3* sont des groupes frères, la distance des paralogues Vertébrés avec leur équivalent protostomien démontre une claire polarité entre les *PTBP* des Vertébrés qui reste difficilement explicable. Nous ne trouvons aucune instance d'un remplacement entre les paralogues, et l'histoire évolutive de chacun d'entre eux est en bonne adéquation avec celle des espèces. L'inadéquation la plus probante entre l'évolution des gènes et des espèces est la polyphylie observée chez les *PTBP1* des mammifères, avec le placement aberrant des monotrèmes et des marsupiaux en dehors de la synapomorphie auquel ils appartiennent, ce qui éclaire déjà sur la spécificité du CUB de ce gène chez les mammifères. Plusieurs résultats vont dans cette direction : i) l'excès d'accumulation des mutations synonymes chez le *PTBP1* des mammifères et ce malgré un taux de mutation similaire (Figure 3.5 B) ; ii) les différences importantes dans le CUB des gènes *PTBP1* de mammifères malgré un nombre total de substitutions similaire (Figure 3.5 A) ; iii) Le spectre unique des mutations synonymes chez *PTBP1*, enrichi en substitutions A->C, T->G et T->C (Figure 3.6) ; iv) La différence significative du CUB entre *PTBP1* et les deux autres paralogues ; v) l'agrégation des gènes *PTBP1* selon leur CUB qui sépare les monotrèmes et les marsupiaux des mammifères pour qu'ils rejoignent le groupe des non-mammifères (Figure3.3 A). D'un point de vue général, la composition nucléotidique et le CUB particuliers des gènes *PTBP1* de mammifères sont probablement associés à des biais mutationnels spécifiques.

Alors que la composition nucléotidique et le CUB GC3-riche des *PTBP1* de mammifères semblent dominés par des biais mutationnels locaux, une telle observation paraît ne pas s'appliquer pour la version *PTBP2* des mêmes organismes. Chez les Vertébrés, la composition nucléotidique varie fortement le long des chromosomes, et ce sous la forme de longues bandes AT-riches ou GC-riches appelées isochores [240]. Les biais mutationnels locaux sont considérés comme l'une des origines principales de ces profils de composition nucléotidique, soulignant alors que la localisation physique d'un gène au sein d'un chromosome façonne grandement son CUB [241]. En accord avec cette hypothèse, les variations dans la composition en GC3 des *PTBP1* sont principalement expliquées par la variation dans la composition nucléotidique des régions non-codantes proximales (Tableau 3.3), suggérant alors qu'un même biais mutationnel a façonné la composition GC-riche des régions flanquantes, introniques et codantes de ce gène. Une même tendance, mais cette fois-ci atténuée, reste vérifiée pour *PTBP3*, mais demeure faible pour *PTBP2*. Pour expliquer un tel enrichissement en AT chez cette version du paralogue *PTBP*, il est nécessaire d'explorer les mécanismes capables d'expliquer un tel CUB, avec peut être un peu plus de profondeur que le biais mutationnel AT-enrichissant que l'on retrouve chez tous les organismes [92, 97, 242].

Chez les mammifères, le profil GC-enrichissant à l'échelle génomique impacte fortement le CUB, et ce de manière à ce que les codons préférentiellement utilisés au sein d'un organisme soient souvent GC-riches [243]. Pour cette raison, le CUB des *PTBP1* de mammifères semble mieux correspondre au CUB global des organismes que les deux autres versions du paralogue. Chez *H. sapiens*, *PTBP1* possède une valeur COUSIN de 1,747, soulignant alors un enrichissement en codons préférentiellement utilisés par l'organisme. Chez *PTBP2-3*, on observe une tendance vers un enrichissement en codons rares, comme l'indiquent les score COUSIN de -0,477 et de -0,235 (Tableau Annexe C.3). La faible complémentarité entre le CUB de *PTBP2* et celui de son organisme pourrait expliquer son expression réduite dans les lignées cellulaires

artificielles humaines et murines, qui sont par ailleurs capables d'exprimer à de forts niveaux *PTBP1* et *PTBP3* [224]. Le principal facteur inhibiteur de l'expression de *PTBP2* semble être relatif à sa traduction. En effet, la modification de la séquence de *PTBP2* pour qu'elle corresponde au CUB de l'organisme conduit à une augmentation spectaculaire de son expresssion, et ce pour des quantités d'ARNm similaires [224]. Une telle stratégie de modification des codons synonymes d'une séquence n'est pas rare, et est même devenu un standard en bio-ingénierie pour augmenter l'expression des gènes, malgré une incompréhension toujours persistante de l'impact et de l'interaction des biais de composition nucléotidiques locaux et globaux sur l'expression d'un gène [244].

L'expression réduite de *PTBP2* dans les cellules humaines et son augmentation par une simple introduction de codons GC-riches en son sein, ainsi que le manque d'impact des biais mutationnels pour expliquer un sa composition nucléotidique conduit à se questionner sur la valeur adaptative de son CUB. La modulation de l'expression de gènes tissu-dépendants ou dépendants du cycle cellulaire a été parfois mentionnée en accord avec le CUB chez *H. sapiens*, comme avec les gènes *TLR7* ou *KRAS* [245–247]. Toujours chez l'Homme, les niveaux d'expression des trois paralogues *PTBP* sont tissu-dépendants (Figure Annexe C.1), et cette différence semble conservée chez les mammifères [248]. Dans le cas d'une duplication de gènes, la sous-fonctionnalisation par le biais d'une spécialisation spatio-temporelle dans l'expression a souvent été proposée comme l'une des force évolutive prédominante de la conservation des gènes paralogues [249]. De tels motifs de régulation spatio-temporelle dans l'expression des paralogues ont été documentés pour un certain nombre de gènes chez une large gamme d'organismes [250–252]. Cette particularité pourrait être expliquée par une présence / absence antagoniste des paralogues au sein d'une même cellule et à un même moment [253]. C'est précisément le cas chez les *PTBP*, où *PTBP1* et *PTBP2* sont antagonistes pendant le développement cérébral : dans les cellules non-cérébrales, *PTBP1* inhibe l'expression de *PTBP2* en sautant l'épissage de son exon 10, alors que durant le développement cérébral, le micro ARN miR124 inhibe l'expression de *PTBP1*, permettant alors la présence de *PTBP2* dans ce tissu [248, 254]. De plus, et malgré le fait que ces deux gènes possèdent tous deux une forte similitude à l'échelle protéique, *PTBP1* et *PTBP2* semblent avoir des activités complémentaires dans la cellule, et affichent une spécificité divergente, définissant alors leur singularité et l'aspect essentiel de leur expression dans certains tissus [255].

Dans une toute autre problématique, nous souhaitons attirer l'attention du lecteur sur la tendance particulière du codon UUG de la Leucine dans notre analyse du CUB. Ce codon est le seul GC3-riche se groupant avec les codons AT3-riches au sein de nos analyses, et ne montre pas d'asymétrie avec son codon antagoniste UUA (Figure 3.3). Un tel comportement pour le codon UUG a été signalé, mais peu discuté, dans d'autres analyses sur le CUB des gènes de mammifères (Figure 7 de l'article de Laurin-Lemay et al. [256]), ainsi que pour les codons AGG et GGG de l'arginine et de la glycine lors de l'étude du CUB au sein d'un grand nombre d'espèces [257]). Les raisons pour lesquelles UUG se regroupe avec les codons enrichis en AT3 demeurent difficiles à déterminer. Une première hypothèse pourrait être fonctionnelle : le codon UUG est particulier, car il peut être un codon alternatif pour démarrer la traduction [258]. Cela dit, d'autres codons tels que les ACG et GUG semblent plus efficaces pour jouer ce rôle, et n'affichent aucune similarité avec les fréquences inhabituelles du codon UUG [259]. Une deuxième ligne de pensée pourrait être relative au répertoire des ARNt, mais les codons UUG et UUA sont tous les deux

décodés par un nombre similaire d'ARNts dans la vaste majorité des génomes (sans compter le « wobble-effect » qui peut s'appliquer sur ces codons) [35]). Au final, une dernière explication suggère que les codons UUG et AGG pourraient être défavorisés si les pressions mutationnelles GC-enrichissantes étaient fortes, et ce malgré le fait que ces deux codons se terminent par un G [35]. En effet, la série de mutations synonymes UUA->UUG->CUG chez la Leucine et la série AGA->AGG->CGG pour l'Arginine souligne l'état transitoire de UUG et AGG pour une augmentation du contenu en GC [35]. Dans notre analyse, AGG se place toutefois avec le reste des codons GC3, en opposition symétrique avec le codon AGA. Le codon UUG se retrouve donc seul à présenter des motifs de fréquences relatives particulières.

Nous avons présenté ici une analyse de l'évolution des gènes paralogues *PTBP*, comme un paradigme de l'évolution sous la duplication d'un gène. Nos résultats montrent que le CUB des *PTBP* a évolué en parallèle d'une expression tissu-spécifique. Dans le cas de *PTBP1*, le plus exprimé au sein des tissus, nous avons identifié le biais mutationnel comme force majeure expliquant le CUB de ce gène. Pour *PTBP2*, l'enrichissement en AT3 pourrait plutôt être compatible avec des forces évolutives associées à une sélection pour une expression spatio-temporelle particulière. Nos résultats suggèrent que l'étude systématique de la composition nucléotidique, de la localisation dans le génome et de l'expression des gènes paralogues peut participer à une meilleure compréhension du complexe mutation-sélection qui façonne le CUB dans les organismes multicellulaires.

II

Partie Deux

ANALYSE DE L'ÉVOLUTION DES POLYOMAVIRUS HUMAINS ET APPLICATIONS AU POLYOMAVIRUS BK DANS LE CADRE DE LA PVAN

Chapitre Quatre

ÉVOLUTION DES POLYOMAVIRUS HUMAINS ET ANALYSE DE LEUR CUB

Ce projet a été réalisé en parallèle des étapes préparatoires pour l'analyse des polyomavirus humains, et plus particulièrement des génomes de BKPyV issus du projet ANR BK-NAB (voir chapitre suivant). Ici, l'objectif est non seulement d'inférer une phylogénie actualisée des polyomavirus humains et des BKPyV, mais aussi de préparer l'analyse des données de patients atteints par la PVAN. En particulier, les phylogénies obtenues au sein de cette étude ont été utilisées pour créer l'outil de génotypage `ViroPhylo`, discuté dans la section suivante. Dans le souci de mieux comprendre la diversité et l'évolution des souches virales de BKPyV, ainsi que de leur CUB, nous nous sommes aussi penchés sur ce type d'analyse. Ce projet a été en partie associé au stage de Master 1 de Luce Faedda (Université de Marseille), que j'ai directement supervisé.

4.1 Matériel et Méthodes

4.1.1 Récupération des données nucléotidiques

Polyomavirus humains

Nous avons récupéré un total de 1434 génomes de polyomavirus humains sur la base de données NCBI [227]. Pour chaque génome, nous avons tenté de récupérer les gènes communs à tous les polyomavirus (*VP1*, *VP2*, *LT* et *ST*), par le biais d'un `blastn` de chaque génome face aux gènes de référence de la souche associée [227]. Les génomes ne possédant pas, à minima, un des gènes cibles ont été écartés du reste de l'analyse. Nous avons vérifié que les séquences extraites étaient bien des CDS et avons donc éliminé de l'analyse toute séquence n'étant pas un multiple de trois et possédant un codon STOP à une autre position qu'à la dernière de la séquence. À la fin de cette étape de préparation, nous avons obtenu un ensemble de données correspondant aux gènes *LT*, *ST*, *VP1* et *VP2* pour les quinze polyomavirus humains. De par le déséquilibre en nombre d'individus entre les polyomavirus considérés au sein de cette étude, nous avons effectué une étape préliminaire de sélection de représentants chez les polyomavirus les plus représentés. Ainsi, nous avons reconstruit la phylogénie en Neighbour-Joining, lorsque c'était possible, de chaque polyomavirus humain. Puis, pour chaque phylogénie, nous avons observé les distances

par paire de chaque individu et avons sélectionné des représentants au sein de chaque groupe observé. C'est donc, dans la mesure où les populations étaient assez nombreuses, un maximum de cinq ou six individus de chaque polyomavirus humain qui ont été gardés pour la suite de l'analyse.

Au final, nous avons obtenu un ensemble de jeu de séquences *VP1*, *VP2*, *LT* et *ST* pour 49 génomes de polyomavirus humains. Un descriptif de ces séquences, sous la concaténation des gènes précoces et tardifs est donnée dans le Tableau 4.1.

BKPyV

L'étape de sélection des génomes complets de BKPyV ne diffère pas de celle décrite précédemment. Un total de 493 génomes de BKPyV ont été sélectionnés pour cette analyse. Contrairement à l'étape précédente, où la diversité génétique des polyomavirus force à une analyse restreinte aux gènes communs, nous avons analysé les BKPyV à l'échelle du génome. Ainsi, nous avons effectué une étape de « rotation » des génomes (voir chapitre suivant) pour que chacun d'entre eux débute sur le premier codon du gène *VP2*, et ce dans le sens 5'-3' des gènes tardifs. Nous avons vérifié que chacun des gènes des BKPyV (*VP1*, *VP2*, *VP3*, *LT*, *ST* et de l'agnoprotéine) soient tous codants et avons supprimé la région NCCR de ces génomes, car bien trop variable pour une analyse phylogénétique. De par la qualité des génomes de BKPyV, c'est un total de 493 génomes, soit le même nombre qu'au départ, qui ont été gardés pour la suite de l'analyse.

4.1.2 Analyse du CUB des polyomavirus humains

Nous avons effectué une analyse du CUB sur les gènes *VP1*, *VP2*, *LT* et *ST* des 49 polyomavirus humains gardés précédemment grâce à l'outil `COUSIN` [94]. Pour chacun d'entre eux, nous avons construit un vecteur de 59 positions contenant les fréquences relatives des codons synonymes, mais avons aussi déterminé leurs scores COUSIN et CAI (Tableau 4.1). Les scores COUSIN et CAI ont été mesurés avec comme référence le CUB du génome humain. Nous avons par la suite effectué une Analyse en Composante Principale (ACP) sur ces mêmes vecteurs de 59 dimensions puis avons effectué des agrégations de données selon la méthode du k-means et du clustering hiérarchique. La Figure 4.1 a été créée par le programme `COUSIN` et fait guise de représentation du CUB des BKPyV.

4.1.3 Alignements et reconstruction phylogénétique des polyomavirus humains et du BKPyV

Afin de générer des alignements robustes sans introduire de quelconques artefacts dus à la forte distance évolutive entre les polyomavirus humains, nous avons effectué plusieurs étapes d'alignement en « cascade » en nous basant sur les résolutions phylogénétiques connues. Ceux-ci ont été réalisés, pour chaque gène, étape par étape à l'échelle protéique :

i Un premier alignement a été effectué pour tous les polyomavirus appartenant à une même espèce.

TABLE 4.1 – **Tableau récapitulatif des valeurs de contenu en GC et de score de CUB des polyomavirus humains**. Ce tabeau est scindé en deux parties distinctes pour les gènes précoces et les gènes tardifs. Au sein de chaque partie, la taille, le contenu en GC, en GC3 et les scores COUSIN et CAI de la concaténation des gènes correspondants sont présentés. Les scores COUSIN et CAI ont été mesurés avec comme référence le CUB du génome humain.

		Région précoce					Région tardive				
Génome (n° d'accès)	**Souche**	**Taille (pb)**	**%GC**	**%GC3**	**COUSIN**	**CAI**	**Taille**	**%GC**	**%GC3**	**COUSIN**	**CAI**
AB211390	HPyV 1	2601	37.14	31.72	-1.00	0.70	2139	42.50	31.42	-0.50	0.70
AB301101	HPyV 1	2595	36.69	31.33	-1.05	0.69	2139	42.50	31.42	-0.55	0.70
AB211369	HPyV 1	2601	36.99	31.03	-1.02	0.70	2139	42.73	31.28	-0.58	0.69
AB211372	HPyV 1	2601	37.14	31.37	-1.03	0.70	2139	42.96	32.12	-0.57	0.69
AB301100	HPyV 1	2601	36.83	30.57	-1.09	0.69	2139	42.45	30.86	-0.62	0.69
AB081611	HPyV 2	2580	38.49	34.07	-0.84	0.72	2094	41.79	30.52	-0.44	0.69
AB126999	HPyV 2	2580	38.30	34.07	-0.86	0.71	2094	41.79	30.52	-0.52	0.69
AF015537	HPyV 2	2580	38.33	33.95	-0.86	0.72	2094	42.07	31.23	-0.49	0.69
AB092581	HPyV 2	2580	38.41	34.19	-0.86	0.71	2094	42.03	31.23	-0.49	0.69
AB127351	HPyV 2	2580	38.92	35.47	-0.80	0.72	2094	42.31	31.23	-0.48	0.69
KU746835	HPyV 3	2496	33.53	26.92	-1.21	0.67	2334	44.82	30.98	-0.32	0.69
KM085447	HPyV 3	2496	33.53	26.92	-1.21	0.67	2334	44.86	31.11	-0.31	0.69
EU358767	HPyV 3	2496	33.41	26.92	-1.21	0.67	2334	44.90	31.23	-0.31	0.69
EU358769	HPyV 4	2526	33.18	24.23	-1.29	0.67	2352	45.37	30.61	-0.19	0.68
KJ725028	HPyV 4	2526	33.02	23.99	-1.30	0.67	2352	45.28	30.74	-0.19	0.69
GU296408	HPyV 4	2526	33.06	23.99	-1.31	0.67	2352	45.45	31.00	-0.15	0.68
EU711058	HPyV 4	2526	32.82	23.40	-1.35	0.67	2352	45.32	30.87	-0.15	0.68
HM011544	HPyV 5	3009	40.38	37.29	-0.31	0.73	1992	40.96	31.63	-0.48	0.68
KF266964	HPyV 5	3009	40.15	36.99	-0.41	0.73	1992	41.06	31.48	-0.46	0.68
HM011540	HPyV 5	3009	40.28	37.19	-0.35	0.73	1992	41.01	31.18	-0.44	0.68
NC_014406	HPyV 6	2577	40.98	37.84	-0.70	0.71	2169	45.64	34.16	-0.36	0.70
MG241567	HPyV 6	2577	41.41	38.53	-0.68	0.71	2169	46.06	34.44	-0.30	0.70
KX379631	HPyV 6	2577	41.25	38.42	-0.66	0.71	2169	46.10	34.58	-0.29	0.70
HM011559	HPyV 6	2577	40.98	37.84	-0.67	0.71	2169	45.74	34.44	-0.33	0.70
HM011564	HPyV 7	2592	40.36	37.62	-0.66	0.71	2127	46.73	37.80	-0.01	0.70
HM011568	HPyV 7	2592	40.32	37.50	-0.65	0.71	2124	46.80	37.29	0.01	0.70
NC_014407	HPyV 7	2592	40.51	37.96	-0.61	0.71	2127	47.02	38.08	0.08	0.71
MN692209	HPyV 7	2592	40.32	37.73	-0.63	0.71	2115	46.86	37.45	0.04	0.70
KX249742	HPyV 8	2685	37.17	33.30	-0.82	0.71	2061	43.09	34.79	-0.37	0.71
KF444093	HPyV 8	2685	37.02	33.07	-0.79	0.71	2061	43.04	34.64	-0.32	0.71
AB873001	HPyV 8	2685	37.10	33.18	-0.80	0.71	2061	43.13	34.79	-0.36	0.71
KF444099	HPyV 8	2685	36.95	32.74	-0.82	0.71	2061	43.04	34.64	-0.32	0.71
NC_015150	HPyV 9	2607	36.13	31.42	-0.70	0.70	2169	42.92	30.71	-0.55	0.68
MH844627	HPyV 9	2607	36.13	31.30	-0.72	0.71	2169	42.92	30.71	-0.55	0.68
HQ696595	HPyV 9	2607	36.13	31.42	-0.70	0.70	2169	42.92	30.71	-0.55	0.68
KC549591	HPyV 10	2601	33.72	24.57	-1.27	0.67	2139	38.99	20.76	-0.82	0.66
KC571705	HPyV 10	2601	33.76	23.76	-1.31	0.67	2139	38.48	19.64	-0.82	0.66
KC571702	HPyV 10	2601	34.03	24.22	-1.25	0.67	2139	39.08	20.90	-0.83	0.66
JQ898291	HPyV 10	2601	33.95	24.34	-1.25	0.67	2139	39.08	21.32	-0.82	0.66
KF530304	HPyV 11	2562	32.87	22.48	-1.35	0.67	2112	38.45	20.03	-0.70	0.65
JX463183	HPyV 11	2562	33.49	24.24	-1.30	0.68	2112	38.78	19.89	-0.64	0.65
KR090571	HPyV 11	2562	33.29	24.24	-1.32	0.67	2112	39.02	20.60	-0.65	0.65
JX308829	HPyV 12	2670	36.74	28.76	-0.97	0.71	2031	43.58	33.38	-0.46	0.67
NC_020890	HPyV 12	2670	36.74	28.76	-0.97	0.71	2031	43.58	33.38	-0.46	0.67
KF954417	HPyV 13	2676	36.51	29.71	-0.97	0.69	2163	42.86	31.76	-0.59	0.69
NC_024118	HPyV 13	2676	36.51	29.71	-0.97	0.69	2163	42.86	31.76	-0.59	0.69
KY404016	HPyV 14	2937	39.80	35.04	-0.59	0.71	2046	40.91	32.99	-0.60	0.71
NC_034253	HPyV 14	2937	39.80	35.04	-0.59	0.71	2046	40.91	32.99	-0.60	0.71
BK010702	HPyV 15	2577	40.36	38.07	-0.64	0.71	2151	46.72	37.10	-0.22	0.70

UUU (F) 1.0	UCU (S) 0.552	UAU (Y) 1.0	UGU (C) 1.0
UUC (F) 0.142	UCC (S) 0.585	UAC (Y) 0.441	UGC (C) 0.538
UUA (L) 1.0	UCA (S) 0.613	UAA (*) 0	UGA (*) 0
UUG (L) 0.697	UCG (S) 0.0	UAG (*) 0	UGG (W) 1.0
CUU (L) 0.731	CCU (P) 1.0	CAU (H) 1.0	CGU (R) 0.033
CUC (L) 0.077	CCC (P) 0.495	CAC (H) 0.393	CGC (R) 0.038
CUA (L) 0.731	CCA (P) 0.765	CAA (Q) 1.0	CGA (R) 0.019
CUG (L) 0.45	CCG (P) 0.0	CAG (Q) 0.553	CGG (R) 0.0
AUU (I) 1.0	ACU (T) 1.0	AAU (N) 1.0	AGU (S) 1.0
AUC (I) 0.032	ACC (T) 0.579	AAC (N) 0.456	AGC (S) 0.299
AUA (I) 0.766	ACA (T) 0.822	AAA (K) 1.0	AGA (R) 1.0
AUG (M) 1.0	ACG (T) 0.0	AAG (K) 0.389	AGG (R) 0.48
GUU (V) 0.794	GCU (A) 1.0	GAU (D) 1.0	GGU (G) 0.605
GUC (V) 0.11	GCC (A) 0.242	GAC (D) 0.451	GGC (G) 0.59
GUA (V) 1.0	GCA (A) 0.273	GAA (E) 1.0	GGA (G) 1.0
GUG (V) 0.494	GCG (A) 0.0	GAG (E) 0.481	GGG (G) 0.668

Figure 4.1 – **Table d'Usage des Codons construite à partir des gènes *LT*, *ST*, *VP1*, *VP2*, *VP3* et de l'agnoprotéine de 493 génomes de BKPyV.** Chaque case donne des informations relatives au codon, à l'acide aminé correspondant (entre parenthèses) et au score RSCU (*i.e* normalisé par la fréquence du codon le plus représenté) associé. Le jeu de couleurs indique l'utilisation du codon par rapport à ses synonymes au sein du jeu de données lu (allant du rouge au vert). Les cases grises indiquent les deux acides aminés codés par un unique codon (tryptophane et méthionine) et les codons STOP (non comptabilisés). Cette Figure est issue d'une analyse avec le programme `COUSIN` [94].

ii Les alignements obtenus ont été combinés avec ceux d'autres polyomavirus selon leur proximité phylogénétique.

iii Ces nouveaux alignements été de nouveaux combinés selon leurs relations phylogénétiques jusqu'à l'obtention d'un unique alignement.

Chacune des étapes d'alignement a été effectuée avec le programme `MAFFT` [260]. Pour la première étape, un alignement classique a été produit et les étapes de combinaison d'alignements ont été produites avec l'aide de la fonction « combine » proposée par `MAFFT` [260]. Tous les alignements obtenus ont été traduits inversement sous leur forme nucléotidique, de manière à ce qu'ils représentent une résolution à l'échelle des codons.

Une fois ces alignements transformés, nous les avons concaténés de manière à obtenir des alignements représentant les gènes précoces (*LT* et *ST*) et les gènes tardifs (*VP1* et *VP2*). Ainsi, les phylogénies découlant de ces différents alignements nous permettent d'explorer l'histoire évolutive des polyomavirus sous plusieurs facettes [173]. Les phylogénies ont été construites sur la base d'un modèle GTR+GAMMA avec une partition par gène et pour chaque position des codons, et ce avec un bootstrap de 1000. Cette étape a été réalisée à l'aide du programme `RAxML` (v8.1) [234].

les génomes de BKPyV ont été alignés avec le programme `MAFFT` sous leur forme nucléotidique avec comme point de départ le premier codon du gène *VP2*. À l'instar des phylogénies décrites précédemment, les phylogénies des BKPyV ont été construites sur la base d'un modèle GTR+GAMMA partitionné sur les gènes et les nucléotides des codons qui composent les séquences, et ce avec un bootstrap de 1000. Cette étape a été réalisée à l'aide du programme `RAxML` (v8.1) [234].

4.2 Résultats

4.2.1 Phylogénie des polyomavirus humains

Nous avons reconstruit la phylogénie des polyomavirus humains à partir de la concaténation des gènes *LT* et *ST* (phylogénie des gènes précoces) et des gènes *VP1* et *VP2* (phylogénie des gènes tardifs). Les deux phylogénies obtenues sont présentées au sein de la Figure 4.2. Ces deux phylogénies regroupent avec une forte solidité les polyomavirus à l'échelle de leur génotype. Le polyomavirus QPyV, considéré comme proche des HPyV6 et HPyV7, forme ici un clade bien supporté avec HPyV7 pour les deux phylogénies[184]. Cela dit, l'absence d'autres QPyV dans la phylogénie ne nous permet pas de trancher avec certitude sur cette position. Les deux phylogénies présentent des différences dans la résolution inter-génotypique de ces virus. La topologie de la phylogénie des polyomavirus à l'échelle des gènes précoces regroupe bien les polyomavirus selon la classification taxonomique de l'ICTV, et nous permet de confirmer l'appartenance du QPyV au clade des *Deltapolyomavirus* [161]. Pour rappel, la classification de l'ICTV se base sur l'analyse du gène *LT* et délimite cinq clades au sein des polyomavirus : *Alphapolyomavirus*, *Betapolyomavirus*, *Deltapolyomavirus* (on retrouve les polyomavirus humains au sein de ces trois clades), *Gammapolyomavirus* et un cinquième groupe sans nom. En contre-partie, la phylogénie

obtenue pour les gènes tardifs propose une résolution plus complexe, où les HPyV3, HPyV4, HPyV6, HPyV7 et QPyV forment un clade à part qui fut un temps considéré comme celui des *Wukipolyomavirus* [173].

4.2.2 Phylogénie du BKPyV

Nous avons reconstruit une phylogénie des BKPyV sur 493 génomes appartenant aux 4 génotypes I, II, III et IV (Figure 4.3 A). La phylogénie obtenue délimite quatre clades distincts rassemblant les individus selon leur génotype et ce avec une définition en parfait accord avec les résolutions réalisées précédemment sur les gènes *LT*, *VP1* et sur leur concaténation. Nous observons toutefois l'insertion d'un unique individu aux abords du clade BK I et dont la position laisse à supposer une événement de recombinaison entre différentes souches de BKPyV. La nature multimodale de la distribution des distances par paires, qui peuvent être utilisées comme guides pour la définition des limites entre catégories, confirme l'existence de certains clades : les BK I et BK IV forment des groupes parmi les plus distants de la phylogénie (Figure 4.3 B). À l'intérieur de chacun de ces génotypes, la phylogénie nous indique une résolution représentant correctement les génotypes de cette échelle taxonomique. Ainsi, les sous-groupes BK a, b-1, b-2 et c sont tous présents et forment des clades distincts, robustes et donc la distance phylogénétique semble confirmer leur existence. Mais il n'en est pas de même au sein des différents sous-groupes de BK IV a-1, a-2, b-1, b-2, c-1, c-2 que l'on pourrait clairement fusionner pour créer les sous-groupes a, b et c. En effet, les individus de ces groupes possèdent une distance similaire aux individus d'un même sous-groupe de BK I. De manière encore plus prononcée, l'observation des individus BK II et BK III, qui forment certes des clades distincts et robustes, pose un questionnement sur l'existence de ces génotypes. En effet, les distances par paires sont plus courtes entre les individus II et III qu'entre les individus des différents sous-groupes de BK I.

4.2.3 Particularités dans le CUB des polyomavirus humains

Nous avons étudié le CUB des polyomavirus humains selon un ACP sur les vecteurs de 59 positions de la concaténation des gènes précoces et tardifs (Figure 4.4). Les deux premiers axes de cette ACP contribuent pour environ 29 % de la variance expliquée. Ces deux axes ne dessinent pas à eux seuls de motif particulier, mais leur combinaison permet d'observer une séparation grossière en diagonale des codons se terminant par AT et par GC. On observe toutefois un rassemblement des codons NCG sur une partie spécifique des deux projections, qui semble évité par la majorité des polyomavirus étudiés (surtout au niveau des gènes tardifs). Le CUB de la concaténation des gènes précoces est, en comparaison avec celui des gènes tardifs, diffus, alors que celui des gènes se concentre majoritairement sur une partie de la projection (à l'exception des gènes tardifs des HPyV 10 et HPyV11), démontrant une légère différence entre les deux groupes de gènes. Si l'on superpose les deux représentations de la projection, on observe une similitude dans le CUPrefs des gènes précoces et tardifs des HPyV10 et HPyV11, mais cette similitude ne semble pas aussi claire au sein des autres HPyV, où la répartition des gènes précoces et tardifs semble différentielle.

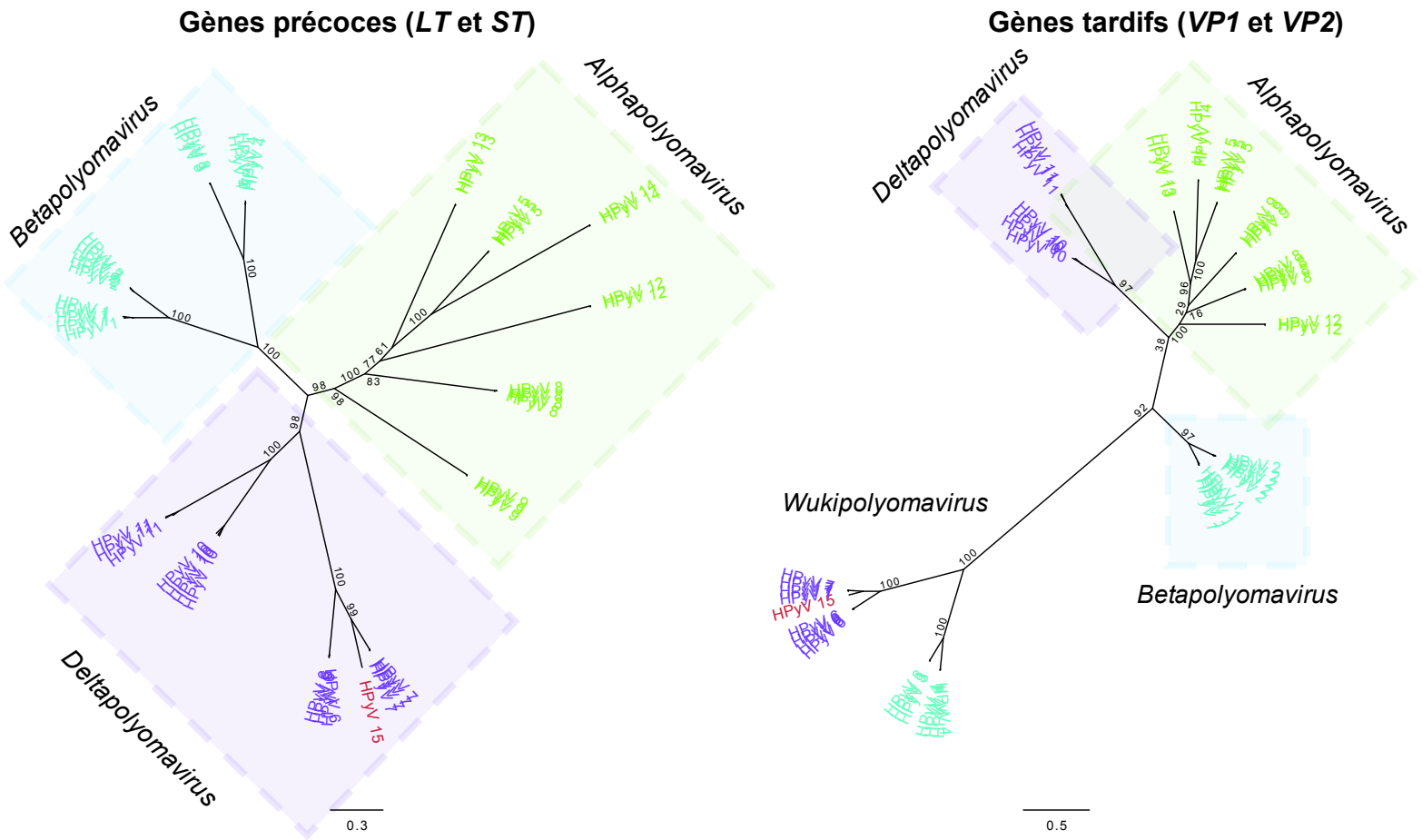

Figure 4.2 – **Phylogénie des gènes précoces (à gauche) et des gènes tardifs (à droite) des polyomavirus humains.** Ces deux phylogénies ont été réalisées à partir de la concaténation des gènes précoces *LT* et *ST* et des gènes tardifs *VP1* et *VP2*, selon une approche en Maximum de Vraisemblance (modèle GTR+GAMMA) avec un bootstrap de 1000 (les valeurs de bootstrap sont indiquées par des chiffres flanquant les noeuds des branches). La phylogénie des gènes précoces regroupe les individus selon la souche à laquelle ils sont rattachés et selon la classification proposée par l'ICTV. On retrouve ainsi les *Alphapolyomavirus* (en vert), *Betapolyomavirus* (en cyan) et les *Deltapolyomavirus* (en violet). La phylogénie des gènes tardifs regroupe elle aussi les individus selon la souche à laquelle ils sont rattachés, mais un clade supplémentaire, anciennement connu sous le nom de *Wukipolyomavirus*, rassemble les HPyV6, HPyV7, QHPyV, WUHPyV et KIHPyV.

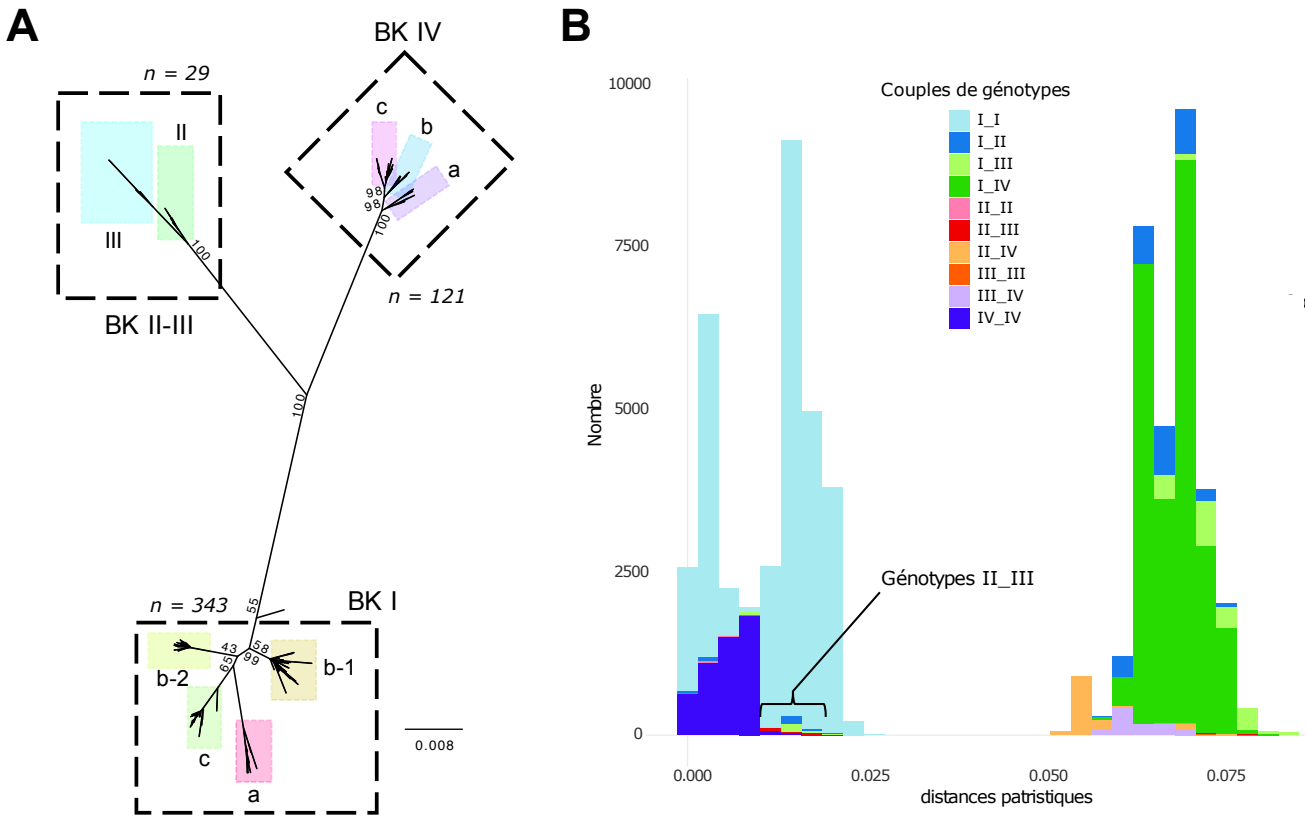

Figure 4.3 – **Phylogénie des BKPyV (A) et distribution des distances patristiques selon le génotype (B).** A) Cette phylogénie a été réalisée à partir des génomes complets (à l'exception de la région NCCR) et selon une approche en Maximum de Vraisemblance (modèle GTR+GAMMA) avec un bootstrap de 1000 (les valeurs de bootstrap sont indiquées par des chiffres flanquant les noeuds des branches). Chaque génotype est délimité par un encadré noir, et les sous-groupes par des encadrés colorés. B) Distribution des distances patristiques observées au sein des différents génotypes de BKPyV. L'axe des abscisses dénote les distances patristiques entre individus et l'axe des ordonnées le nombre de paires d'individus possédant une distance patristique similaire. Les couleurs de cet histogramme indiquent les génotypes auxquels sont rattachés les couples d'individus (voir légende). On observe trois modes au sein de cet histogramme. Les deux premiers modes rassemblent majoritairement les individus appartenant à un même sous-groupe et à un même génotype, alors que le troisième mode, éloigné des deux autres, rassemble les individus appartenant à différents génotypes. Au sein de cette distribution, on observe que les distances patristiques des individus BK II et BK III se rassemblent au même niveau que les individus intra-génotypes de BK I (et accessoirement de BK IV). De manière similaire, les relations inter-sous-groupes des BK IV sont majoritairement condensées au même niveau que les relations intra-groupes de la majorité des BK I.

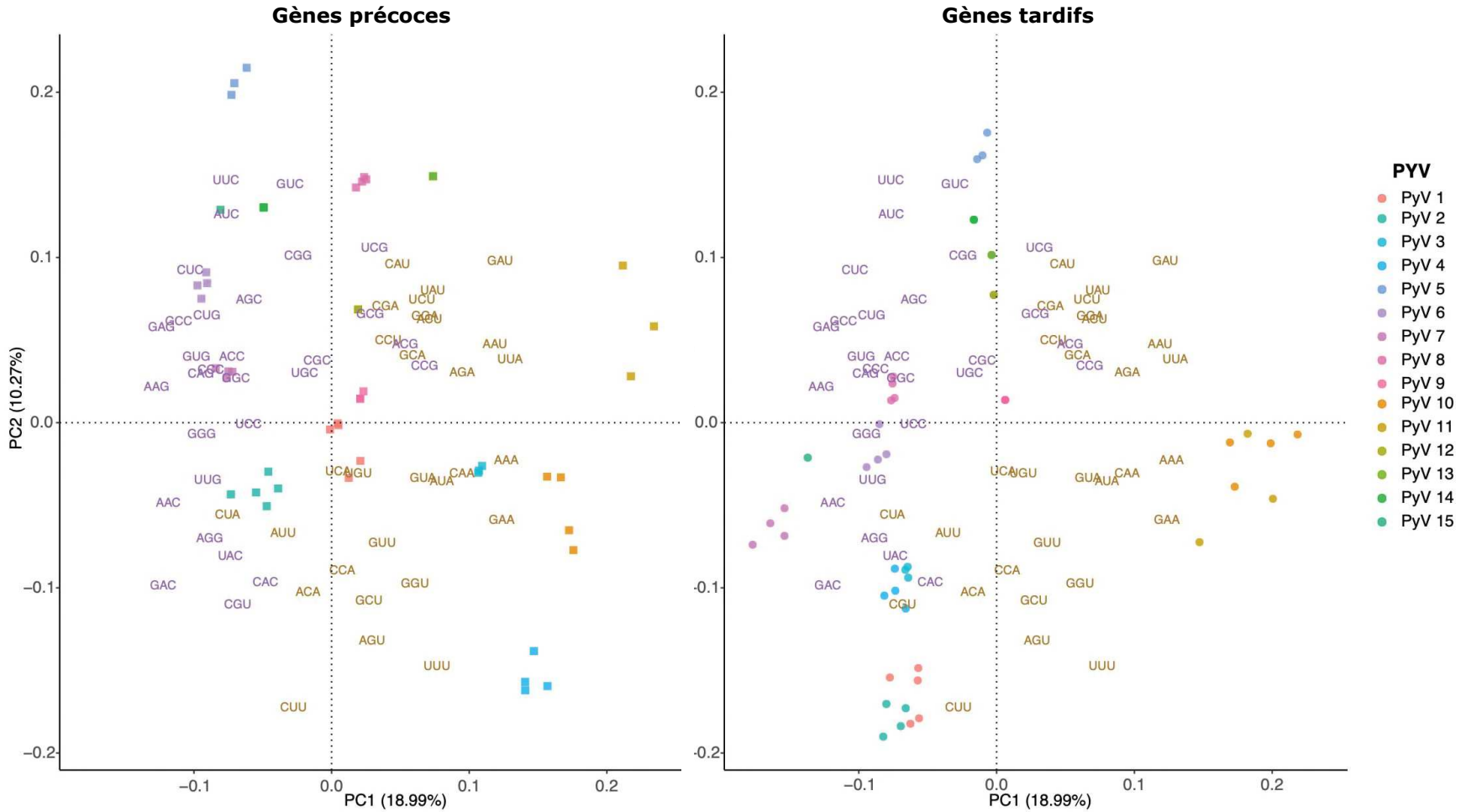

Figure 4.4 – **Projection des deux premiers axes d'une Analyse en Composante Principale sur le CUB des gènes précoces (A) et tardifs (B) des polyomavirus humains.** Chacun des deux graphes correspond à la même ACP. Celui de gauche indique la dispersion des gènes précoces sur les deux premiers axes de l'ACP et celui de droite la dispersion des gènes tardifs.

Nous avons par la suite cherché à déterminer la relation entre le CUB des gènes précoces et tardifs et leur histoire évolutive. Nous avons pour cela mesuré les distances patristiques entre génomes sur les arbres phylogénetiques construits pour les gènes précoces et tardifs, et avons mesuré les distances euclidiennes sur les vecteurs des fréquences de 59 codons synonymes des gènes. La comparaison entre les distances euclidiennes du CUB et la distance patristique des individus au sein des phylogénies précoces ou tardives indique une corrélation positive d'environ 0.60 entre les deux variables, et ce quelle que soit la nature des gènes observés (Figure 4.5). Mais ces résultats pourraient être faussés par la proximité de la relation entre les individus appartenant à un même génotype. En effet, lorsque l'on observe la résolution des dendrogrammes issus des clustering hiérarchiques pour les gènes précoces et tardifs, on observe un désagrément total avec leurs phylogénies respectives. Seule la résolution intra-groupe est respectée, alors que la topologie observée n'est pas proche de la classification de l'ICTV. On pourra toutefois observer le regroupement de certains polyomavirus, comme les HPyV3 et HPyV4 pour les gènes précoces et des HPyV6 et HPyV7 pour les gènes tardifs. Pour finir, ces deux dendrogrammes ne sont pas en accord, et présentent des différences dans la topologie globale des HPyV, signifiant alors une claire discordance dans le CUB des gènes précoces et des gènes tardifs (Figure 4.6). Mais il est important de pondérer ces résultats par le fait que ces différences pourraient être dues au désaccord des histoires évolutives pour les gènes précoces et tardifs.

4.3 Discussion

Ce projet préliminaire a pour objectif de faire état de l'avancement de nos connaissances sur l'histoire évolutive des polyomavirus humains et de leur CUB, mais aussi de construire notre propre jeu de données pour préparer l'analyse de polyomavirus humains et en particulier des BKPyV.

Au sein de cette analyse, nous avons pu confirmer la position du polyomavirus QPyV récemment isolé comme étant proche des polyomavirus HPyV6 et HPyV7, où QPyV forme un clade robuste avec HPyV7. Ce polyomavirus a été isolé lors d'une analyse de métagénomique virale sur des échantillons de selles, et a été considéré comme souche à part entière de par sa distance nucléotidique avec les HPyV6 et HPyV7 [184]. Bien qu'il soit nécessaire d'isoler de nouveaux QPyV pour confirmer l'existence de son génotype, nous pouvons considérer que celui-ci appartiendrait au clade des *Deltapolyomavirus*. Comme attendu, les résolutions phylogénétiques basées sur les gènes précoces et les gènes tardifs sont divergentes. L'existence d'un clade *Wukipolyomavirus* au sein de la phylogénie des gènes tardifs est en parfaite adéquation avec les observations faites par *Buck et al.*. Ce clade serait apparu suite à des événements de recombinaisons au sein de certains polyomavirus humains tels que les HPyV3, HPyV4 ou encore le HPyV6. Ces événements de recombinaison n'auraient eu lieu qu'au niveau des gènes tardifs, alors que les gènes précoces auraient été soumis à une évolution plus « linéaire » [173]. De par le placement phylogénétique du QPyV au sein des deux phylogénies, toujours à proximité du HPyV7, on pourrait considérer que ce polyomavirus a une histoire évolutive proche de ses voisins, aussi bien au niveau des gènes précoces que des gènes tardifs. Il aurait donc subi les mêmes événements de recombinaison que les autres *Wukipolyomavirus*.

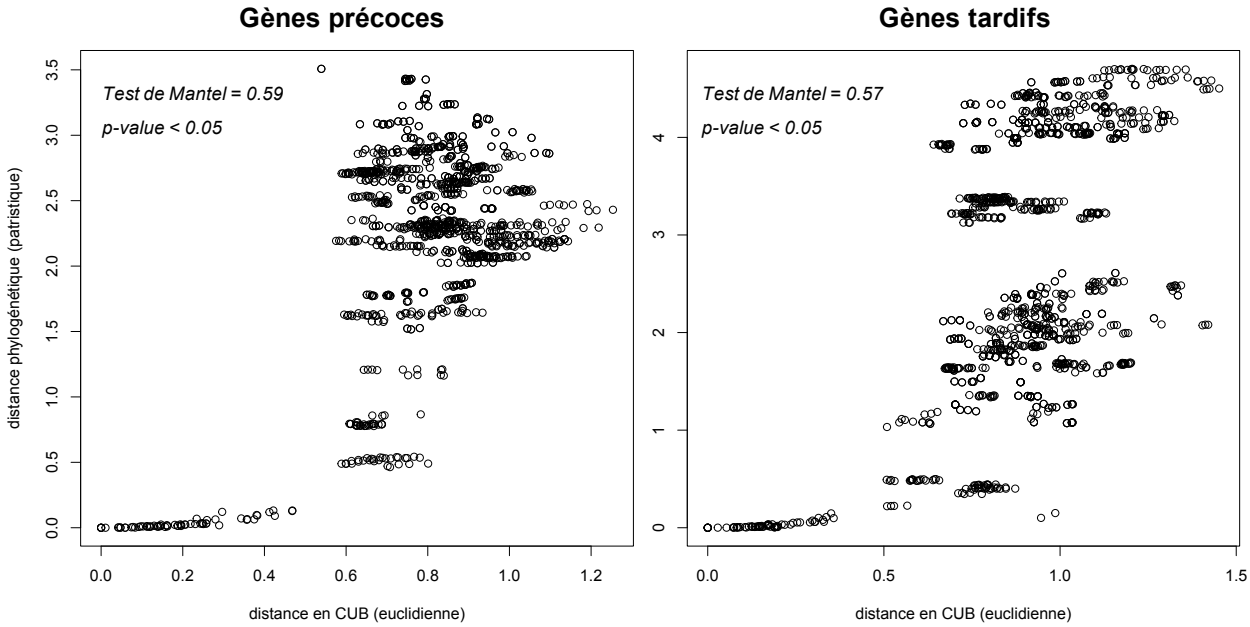

Figure 4.5 – **Évaluation de la concordance entre distances patristiques des phylogénies et distances euclidiennes du CUB des gènes précoces et tardifs.** Les matrices ont été obtenues à partir des phylogénies des gènes précoces et tardifs, et des fréquences des 59 codons synonymes. La concordance a été mesurée à partir de tests de Mantel dont les résultats principaux sont affichés au sein des deux sous figures.

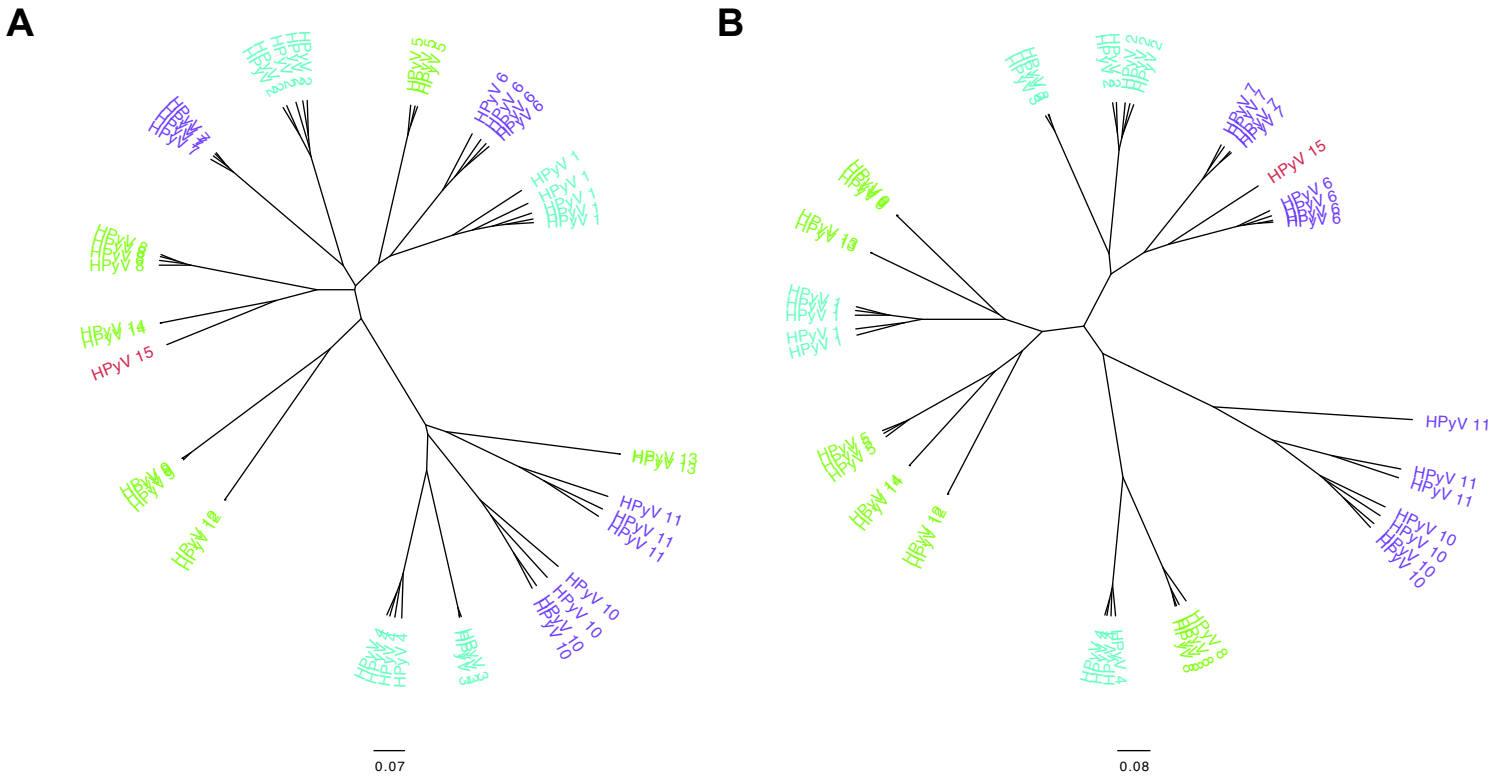

Figure 4.6 – **Dendrogrammes des gènes précoces et des gènes tardifs obtenus à la suite d'un clustering hiérarchique sur le CUB de ces gènes.**. Chaque individu est indentifié par la souche auquel il appartient, et le jeu de couleurs indique les clades auxquels ils appartiennent (selon la classification de l'ICTV basée sur le gène *LT*) [161].

L'histoire évolutive des BKPyV observée au sein de cette étude est conforme aux phylogénies précédemment produites. Cela dit, nous avons pu observer un certain manque de consistance quant à la classification proposée ultérieurement. La distance phylogénétique des sous-groupes du BKPyV I est hautement plus grande que celle retrouvée au sein des sous-groupes du BKPyV IV, au point où l'on pourrait fusionner les six sous-groupes de BK IV en trois sous-groupes a, b et c. Mais l'observation est encore plus marquée au niveau des BK II et BK III. *Krumbholz et al.* avaient déjà proposé au sein de leurs travaux que les BK II et BK III formeraient en réalité un unique clade, que l'on peut considérer comme à la même échelle taxonomique que les sous-groupes BK I au sein de notre propre étude [199]. Ici, nous confirmons cette tendance et avertissons le lecteur sur l'interprétation des génotypes BK II et BK III. Ceux-ci ne devraient pas être considérés à la même échelle que les BK I et BK IV. Nous observons par ailleurs l'insertion d'un individu aux abords du groupe BK I, dont la position pourrait bel et bien indiquer un événement de recombinaison au sein du génome de cet individu.

Nous avons aussi remarqué que le lien entre l'évolution des polyomavirus et celui de leur CUB n'est pas clair. Bien que les génomes d'un même génotype arborent, naturellement, un CUB similaire, il n'en est pas de même à une échelle taxonomique plus haute. Par rapport aux deux phylogénies discutées auparavant, nous observons une résolution tout à fait différente au niveau du clustering hiérarchique de ces mêmes groupes de gènes. Il en est de même lorsque l'on compare le CUB des gènes précoces et des gènes tardifs. Il est intéressant de noter que la grande majorité des gènes précoces et tardifs possèdent un biais d'usage des codons contraire à celui de leur hôte humain (voir scores COUSIN de la Table 4.1). En suivant les hypothèses de *Chen et al.*, on pourrait supposer une « désoptimisation » du CUB des polyomavirus humains pour que ceux-ci ne provoquent pas d'infection symptomatique [131]. Mais il existe une différence dans les scores COUSIN des gènes précoces et tardifs : les gènes tardifs possèdent un CUB plus proche d'un usage équiprobable des codons et, à une mesure plus discutable, du CUB de l'hôte humain que les gènes précoces. Ces résultats se rapprochent de ceux observés pour les papillomavirus humains, où le CUB des gènes tardifs est plus proche de celui de l'hôte que ne l'est le CUB des gènes précoces [129]. Comme les gènes précoces sont moins exprimés que les gènes tardifs chez les papillomavirus, on pourrait supposer que la différence dans le CUB suivrait les hypothèses d'atténuation de l'expression par le CUB de *Chen et al.*, et que ce mécanisme pourrait être aussi exploité par les polyomavirus humains [131]. Bien entendu, de telles spéculations devront être confirmées par une mesure des niveaux de protéines virales au cours d'une infection par un polyomavirus. Mais il est toutefois nécessaire de remarquer que la grande majorité des HPyV possèdent un CUB similaire, et que les différences observées relèvent peut être de signatures négligeables. La Figure 4.1, qui définit le CUB des BKPyV de cette étude, fait état d'une utilisation quasi-nulle des codons possédant un CpG, pourvu que d'autres synonymes ne contenant pas ce dinucléotide existent. Cette tendance semble vérifiée chez les autres polyomavirus, où les codon NCG semblent évités (Figure 4.4). Ainsi, on observe par exemple au sein des codons de l'arginine un évitement systématique des NCG, au profit des codons AGA et AGG. Un tel évitement est probablement issu d'un mécanisme de protection du matériel génétique viral contre l'action de protéines tel que les *ZAP* et les *Toll-like receptors* [85, 132].

Les résultats obtenus au sein de cette analyse sont préliminaires, et les conclusions que nous pouvons en tirer sont clairement spéculatives. Il est nécessaire d'approfondir notre analyse, en

prenant par exemple l'étude des *PTBP* comme source d'inspiration pour déterminer la nature du CUB et ce qu'il implique chez les polyomavirus humains. Pour aller plus loin dans l'analyse du CUB de ces virus, l'une des pistes les plus intéressantes serait de nous focaliser sur plusieurs métadonnées pour vérifier l'existence de liens entre CUB et tropisme cellulaire, répartition géographique des HPyV, sévérité dans le cadre d'une infection symptomatique ou encore et tout simplement dans la charge virale observée. L'observation de tendances (ou non) démontrant l'existence d'un lien entre ces variables permettrait alors comprendre le rôle putatif du CUB dans l'expression des gènes viraux et dans la relation hôte-parasite.

Chapitre Cinq

5

ANALYSER LES BKPYV ET HUMAINS AU TRAVERS DE DEUX PIPELINES ET D'UN MODÈLE MATHÉMATIQUE

5.1 Introduction

Le séquençage de l'ADN à Haut-Débit (SHD ; HTS en anglais pour *High-throughput Sequencing*, aussi appelé NGS pour *Next-Generation Sequencing*) consiste en la détermination de la séquence nucléotidique d'une région génomique, d'un génome ou encore d'un métagénome à partir d'un échantillon donné et selon des stratégies permettant un séquençage rapide et conséquent [261]. Pour ce faire, la majeure partie des différentes technologies de séquençage se basent sur le principe de polymérisation étape par étape de la méthode de Sanger [262] et fonctionnent, de manière non-exhaustive, de la sorte :

i Extraction de l'ADN de l'échantillon. Selon la nature de l'échantillon et la quantité d'ADN qu'il contient, cette étape peut être suivie par une pré-amplification du matériel génétique pour augmenter le nombre de copies d'ADN/mL. [261, 263].

ii Cassure de l'ADN en fragments dans des tailles différentes en fonction de la méthode de cassure décrite en aval, et préparation de « librairies » (ou banque de fragments) en vue du séquençage. La cassure de l'ADN peut être faite via sonication, nébulisation ou encore par le biais d'une activité enzymatique. La plupart du temps, les fragments sont produits « à l'aveugle », c'est à dire qu'il n'y a pas de site de cassure prédéfini. La construction de la librairie se fait par l'ajout de courtes séquences, appelées « adaptateurs » sur les fragments d'ADN, de manière à pouvoir les identifier (*barcoding*) et à permettre leur fixation sur le substrat du séquenceur (*e.g.* une plaque appelée *flow-cell* contenant des séquences complémentaires des adaptateurs). [263]

iii Dénaturation des fragments, fixation sur le substrat, PCR et séquençage en temps réel des nucléotides par polymérisation du brin complémentaire des fragments de PCR sur le substrat du séquenceur. Généralement, le séquençage se fait étape par étape par ajouts successifs de milieux ne contenant qu'un seul nucléotide. Si ce nucléotide est ajouté à un fragment

d'ADN en cours de réplication, celui-ci est analysé par le séquenceur pour confirmer (ou infirmer) la polymérisation. La lecture successive des nucléotides insérés lors de la réplication permet alors de connaître la séquence d'ADN associée au fragment. Mais il existe d'autres approches, comme la lecture directe du fragment d'ADN (sans passer par une quelconque polymérisation) avec l'approche Nanopore [264]. Les séquences analysées sont appelées *reads* (lecture en français). [265]

iv Assemblage bioinformatique des *reads* par méthodes dites de *mapping* (à partir d'un génome de référence) ou *de novo* (par assemblage des *reads* sans génome de référence). [266–268].

On peut considérer qu'il existe deux grandes approches pour séquencer l'ADN : celles qui se basent sur le séquençage de *short reads* (où les *reads* ne dépassent pas quelques centaines de paires de bases) [261, 269] et celles qui proposent un séquençage sur de *long reads* (dont la taille peut atteindre plusieurs milliers de paires de bases) [261, 270–272]. Les approches *short reads* possèdent l'avantage d'être plus précises et peu coûteuses, mais l'assemblage du génome ou de la région génomique par association des petites lectures peut être fastidieux et, dans certains cas, mener à des erreurs dans l'agencement de l'ADN [261]. Jusqu'à aujourd'hui, les approches *long reads* étaient bien trop peu précises (et le demeurent toujours à un certain niveau), mais leur principal avantage est qu'elles représentent avec une meilleure précision l'organisation de la molécule assemblée [261]. La méthode de séquençage PacBio (*Pacific Biosciences*) CCS (*Circular Consensus Sequencing* ou séquençage consensus circulaire en français), que nous décrivons ici car utilisée pour séquencer les BKPyV que nous étudierons, est une technologie de séquençage *long reads* dont la précision est proche d'une méthode *short reads*. Le séquençage CCS se base sur le concept SMRT (*Single Molecule Real-time Sequencing* ou séquençage en temps réel à partir d'une seule molécule en français) [261, 272]. Il fonctionne de la sorte :

i Les fragments d'ADN sont isolés, linéarisés si besoin et sont rattachés à des adaptateurs de part et d'autre de leur molécule. Ces adaptateurs forment une boucle de chaque côté du fragment d'ADN, si bien que le fragment est « circularisé ». Une ADN polymérase est rattachée au fragment d'ADN sur l'un des adaptateurs.

ii Les fragments modifiés sont placés sur une plaque possédant des puits appelés ZMW (*Zero mode waveguides*). L'activité enzymatique de la polymérase débute une fois le fragment d'ADN placé dans un puits, et l'insertion de chaque nouveau nucléotide lors de cette réplication émet une fluorescence particulière qui est captée par le système ZMW.

iii La particularité du séquençage CCS est que les fragments circularisés sont répliqués un certain nombre fois par la polymérase. De ce fait, le fragment est analysé plusieurs fois, et le *read* obtenu (appelé *Hifi*) est forme une séquence consensus de l'ADN fragmenté. De par la lecture répétée du fragment d'ADN et de la formation d'une séquence consensus individuelle pour chaque molécule, le séquençage CCS corrige les erreurs classiquement retrouvées au sein des méthodes de séquençage *long read* [261, 272].

La métagénomique virale se base, à titre d'exemple, sur l'analyse du matériel génétique viral retrouvé au sein d'un échantillon de sang ou d'urine [273]. Ces analyses de métagénomique peuvent viser un spectre large (*i.e.* représenter la diversité virale totale de l'échantillon) ou plus

réduit (*i.e.*, rechercher la présence et la diversité de certains virus au sein de l'échantillon). Dans le premier comme dans le dernier cas, il est parfois difficile de discriminer l'ADN viral d'une souche des autres ADN viraux, de l'ADN bactérien ou encore de l'ADN eucaryote que l'on peut retrouver au sein de l'échantillon [274]. Sans a priori sur le virome analysé, il est nécessaire de discriminer les populations individuelles par une inférence statistique des OTU (Operational Taxonomic Unity ; unité taxonomique opérationnelle en français), qui se définit par une mesure de la distance génétique des génomes ou fragments de génomes assemblés [275]. Avec a priori sur le virome analysé, l'approche demeure différente : on cherche à déterminer la quantité et la diversité virale des souches attendues, notamment par une détermination qualitative et quantitative des *reads* lors d'un mapping. C'est sur ce dernier point que ce projet trouve sa source.

Lors de l'analyse ciblée de la métagénomique virale des BKPyV chez un patient immunosupprimé, l'un des défis est de discriminer et de représenter avec précision les populations virales. Un séquençage par une approche de short read peut conduire, lors de l'assemblage des génomes, à la reconstruction de génomes hybrides qui ne représenteront pas la diversité génotypique de l'échantillon, car trop peu distincts pour être discriminés. Il est donc nécessaire de privilégier les approches *longs reads*, si possible pour séquencer un génome entier, tout en gardant à l'esprit que les erreurs générées lors du séquençage pourront être corrigées par une approche *short read* supplémentaire sur les mêmes échantillons. C'est pour cette raison que les échantillons de l'ANR BK-NAB ont étés analysés à partir d'une technologie CCS, car celle-ci permet d'analyser des génomes complets sans passer par l'étape de fragmentation de l'ADN. Chaque molécule analysée par le séquenceur correspond donc à un génome complet.

Nous avons développé deux pipelines d'analyse de génomes complets de BKPyV appelés `GenoPolys` et `ViroPhylo`. Ceux-ci ont pour objectif de nettoyer et de préparer les séquences génomiques issues des séquençages CCS PacBio pour de futures analyses. Ces séquences ont été obtenues à partir des échantillons longitudinaux d'urine et de sang de patients receveurs d'une greffe de rein (souffrant ou non d'une PVAN). D'autre part, nous proposons une révision du modèle mathématique d'évolution intra-hôte des BKPyV de *Funk et al.*, en vue d'une analyse de la diversité virale et de l'évolution de la virémie et de la virurie chez ces mêmes patients [1]

5.2 Pipeline `GenoPolys`

5.2.1 Architecture et fonctionnement du pipeline

Le pipeline `GenoPolys` a été construit pour traiter des données provenant d'un séquençage CCS de BKPyV. Cela dit, ce pipeline peut parfaitement s'adapter aux données issues de n'importe quelle autre technologie de séquençage. Ce traitement de données passe par une étape obligatoire de sélection de données fiables, génotype les polyomavirus sélectionnés et propose par la suite plusieurs étapes permettant l'obtention de données sur les gènes *LT*, *ST*, *VP1*, *VP2*, *VP3*, de l'agnoprotéine et de la région NCCR. L'architecture de ce pipeline est décrite au sein de la Figure 5.1.

Dans un premier temps, les données brutes issues du séquençage sont nettoyées et sélectionnées pour la suite de l'analyse. Une détection d'artefacts au sein des séquences issues d'un

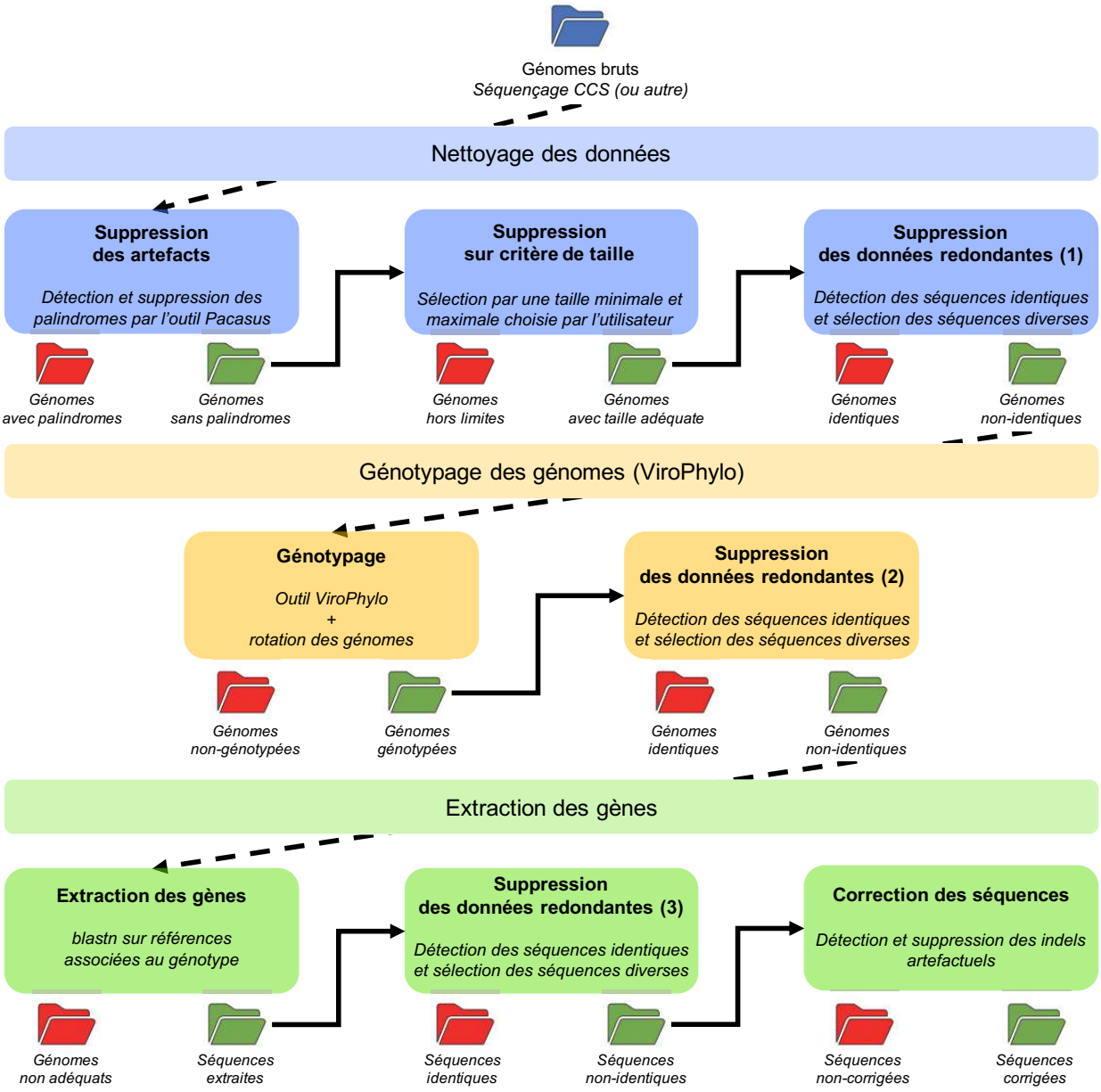

Figure 5.1 – **Pipeline de l'outil** `GenoPolys`**.** Les données génomiques insérées par l'utilisateur subissent tout d'abord une étape dite de nettoyage des données. Une vérification de l'existence de palindromes est effectuée par l'outil `Pacasus`. Les génomes ne contenant pas de palindromes sont ensuite discriminés par un critère de taille, et ce pour empêcher tout autre artefact non-palindromique d'importance. Puis, une étape de vérification sur l'existence de génomes identiques est effectuée. Ceux qui présentent une identité exacte sont mis de côté, mais sont à considérer car représentatifs de la quantité de génomes similaires analysés. Les génomes sélectionnés pour la suite du programme sont ensuite analysés par l'outil `ViroPhylo`, qui permet non seulement de les génotyper, mais aussi de les uniformiser pour que chaque génome débute au premier codon du gène *VP2*. Une fois les génomes identifiés, ceux-ci subissent une étape d'extraction des gènes et de la NCCR par *blast* des génomes sur des gènes de références associés à leur génotype. Les gènes extraits subissent une nouvelle étape de délétion des identiques et sont ensuite corrigés pour éviter tout biais de séquençage lié au indels artefactuels sur les homopolymères.

séquençage CCS de PacBio est effectuée à l'aide de l'outil `Pacasus` [276]. Celui-ci a été spécialement conçu pour détecter et pour supprimer les palindromes artefactuels issus de l'étape de pré-amplification d'un séquençage *long reads* [276]. Ces palindromes artefactuels sont définis comme la répétition inversée d'une région du fragment d'ADN, ce qui provoque l'apparition de régions hybrides au sein du génome, faussant alors son contenu. `Pacasus` détecte, avec une certaine liberté relative au taux d'erreur des méthodes *long read*, les palindromes et les supprime de l'analyse [276]. Il existe des variations dans la taille et dans le contenu des séquences NCCR, et certaines d'entre elles peuvent conduire à l'apparition de « palindromes » naturels qui, techniquement, pourraient être supprimés de l'analyse par `Pacasus`. Nous conseillons fortement aux utilisateurs de vérifier le contenu des séquences supprimées pour vérifier la fiabilité de l'étape de suppression des palindromes. Les données conservées à la fin de cette première analyse sont ensuite discriminées selon leur taille. Pour la raison citée ci-dessus, la NCCR est sujette à des variations de contenu pouvant modifier sa taille, mais certains artefacts non-détectés par `Pacasus` peuvent persister, et il est donc nécessaire de les discriminer. Dès lors que la taille du génome est trop petite (ou trop grande), il est tout à fait acceptable de mettre ces données de côté, quitte à analyser leurs particularités par la suite. Le choix de la taille minimale et maximale acceptée est au choix de l'utilisateur. Pour finir, les séquences identiques sont supprimées, car l'objectif primaire de cette analyse est de décrire la diversité virale au sein d'un échantillon. Un décompte du nombre de séquences identiques est toutefois effectué, car cette répétition représente une approximation des différences en fréquences des populations virales retrouvées au sein de l'échantillon. Une fois les données préparées, les génomes sont génotypés à l'aide de l'outil *standalone* `ViroPhylo` (discuté dans la section suivante). Au cours du génotypage, les génomes font l'objet d'une « rotation » consistant en une transformation des séquences pour que celles-ci débutent au premier codon du gène *VP2*. Cette uniformisation prépare les génomes pour de futures étapes de traitement des données. À partir de ce nouveau jeu de données, les différents gènes et la région NCCR sont extraits des génomes par alignement sur des séquences de référence du génotype correspondant, et ce à l'aide de la fonction `blastn` du logiciel `BLAST+` [227]. Cette étape est suivie par une suppression des séquences identiques afin de déterminer la diversité nucléotidique au sein des gènes et de la région NCCR. Encore une fois, un décompte des séquences extraites identiques est effectué, et ce pour déterminer les différences en fréquence des gènes et de la région NCCR. Les séquences extraites sont ensuite corrigées via alignement face à une référence du CDS correspondant. Cette dernière étape repose sur la correction d'erreurs de séquençage CCS qui provoque l'apparition d'indels artefactuels, généralement au sein d'un homopolymère [277]. Un indel provoque un décalage du cadre de lecture rendant la protéine associée au CDS non-fonctionnelle. La correction vise alors à supprimer le nucléotide artefactuel ou à rajouter un nucléotide « N » en fonction de la nature de l'indel. En effet, même s'il a été démontré que les populations de virus peuvent contenir naturellement des CDS ne pouvant donner des protéines viables [278–280], l'objectif de ce pipeline est de déterminer la diversité au sein d'une population de virus effective, et le taux important d'indels artefactuels nous oblige à sous-évaluer cette diversité par correction des séquences.

Au final, le pipeline `GenoPolys` donne en sortie les résultats obtenus de manière à ce que l'utilisateur ne perde aucune donnée, mais plutôt que celles-ci soient hiérarchisées selon leur sélection ou leur délétion au cours du déroulement du programme.

5.2.2 Fonctionnement du pipeline `GenoPolys` sur des données préliminaires de l'ANR BK-NAB

Le pipeline `GenoPolys` permet de traiter les génomes de BKPyV pour les uniformiser, dénombrer leur quantité et leur diversité, mais aussi pour isoler les gènes et la région NCCR pour de futures analyses. Pour le tester, `GenoPolys` a été utilisé sur des données préliminaires issues d'un séquençage CCS (Tableau 5.1). Les échantillons d'urine de neuf patients ont étés analysés de manière longitudinale. Chaque patient possède le même identifiant, et les différents prélèvements sont indiqués par un nombre suivant l'identifiant, comme c'est le cas chez $LAGU\_4$ et $LAGU\_28$. Lors de l'étape de nettoyage de données brutes, peu de séquences sont supprimées de l'analyse par `Pacasus` et la discrimination par taille, à l'exception des échantillons $BROYV\_35-1$ et $BROYV\_40-1$ pour lesquels nous avons dû supprimer un grand nombre de génomes possédant des artefacts palindromiques (seuls 67 % et 20 % des génomes ont été gardés après analyse par le logiciel de `Pacasus`). De tels résultats pour ces deux échantillons pourraient être expliqués par une forte prévalence de palindromes au sein des génomes issus de l'étape de pré-amplification. En effet, le faible nombre initial de génomes séquencés a conduit à plusieurs étapes de pré-amplification de l'ADN pour qu'il soit séquencé convenablement, qui ont inévitablement conduit à l'apparition d'un grand nombre de palindromes. Un tel nombre de génomes initiaux dans ces échantillons pourrait, à titre d'exemple, s'expliquer par des biais expérimentaux lors de l'extraction de l'ADN, ou tout simplement par une faible virurie au moment du prélèvement. Tous les génomes obtenus à la fin de ces deux étapes de sélection de données différaient, démontrant une forte diversité au sein des différents échantillons. Une fois les génomes génotypés et transformés pour qu'ils débutent sur le gène *VP2*, une extraction des gènes et de la région NCCR a été effectuée. Celle-ci a, pour chaque échantillon, extrait avec succès chaque gène et la NCCR. En contre-partie, l'étape de suppression des données redondantes n'a pas donné les mêmes résultats pour chaque échantillon. En général, le gène *LT* semble diversifié au sein des séquences obtenues, mais la diversité est bien plus basse au sein des gènes *VP1*, *VP2*, *ST*, de l'agnoprotéine et de la région non-codante NNCR. À titre d'exemple, certains échantillons tels que $LAGU\_4$ possèdent une forte diversité au sein du gène *LT* (100 % de gènes différant sur au moins un nucléotide) et un taux plus nuancé au niveau des gènes *VP1* (17 %), *VP2* (48 %), *ST* (24 %), de l'agnoprotéine (40 %) et de la région non-codante NCCR (46 %). Il est évident que les données obtenues à la fin d'une analyse `GenoPolys` ne sont que préliminaires, et permettent, en plus de ses tâches de traitement des données génomiques, de faire un simple état de la diversité des génotypes, des génomes et des gènes observés. Pour poursuivre une telle analyse, il est nécessaire d'effectuer de réelles déterminations de la diversité génétique de manière longitudinale au sein des échantillons d'un même patient, et de poursuivre cet effort par des mesures telle que celle de la diversité nucléotidique π . *In fine*, GenoPolys pourra répondre aux besoins d'une méthode fiable d'étude de génomique des populations intra-hôte des BKPyV.

TABLE 5.1 – **Résultats obtenus à la suite d'une analyse `GenoPolys` sur des génomes issus d'échantillons d'urine de patients receveurs d'une greffe de rein.** Les pseudonymes des individus analysés sont indiqués dans la colonne « Échantillon » . Certains échantillons possèdent un même identifiant, mais diffèrent dans le numéro attribué, indiquant alors un échantillonnage sur un même patient à deux différentes dates. Les colonnes « `Pacasus` », « Taille » et « Total » indiquent le nombre de génomes conservés (ainsi que leur proportion par rapport à l'étape précédente) après chaque étape de l'analyse de nettoyage des données. Les colonnes « *LT* », « *ST* », « *VP1* », « *VP2* », « *VP3* », « agnoprot. » (pour agnoprotéine) et « NCCR » indiquent le nombre de séquences retrouvées à la suite de l'extraction des gènes et de la suppression des gènes identiques (il est à noter que l'étape d'extraction a pu extraire toutes les régions d'intérêt, quel que soit le génome analysé). Ces dernières colonnes font donc état de la diversité des gènes et de la région NCCR au sein des échantillons obtenus.

		Nettoyage des données			**Extraction des gènes et de la NCCR (après délétion des séquences identiques)**						
Échantillon	**Génomes**	`Pacasus`	**Taille**	**total**	***LT***	**ST**	***VP1***	***VP2***	***VP3***	**agnoprot.**	**NCCR**
VQQ	6472	6272 (0.97)	6159 (0.98)	6159 (0.95)	5233 (0.85)	1508 (0.24)	6155 (1)	2931 (0.48)	2401 (0.39)	2445 (0.40)	2829 (0.46)
LAGU_4	4421	4261 (0.96)	4197 (0.98)	4197 (0.95)	4194 (1)	1313 (0.31)	734 (0.17)	2036 (0.49)	1721 (0.41)	1691 (0.40)	4195 (1)
LAGU_28	15548	15180 (0.98)	14864 (0.98)	14864 (0.96)	11383 (0.77)	4906 (0.33)	14852 (1)	14862 (1)	14615 (0.98)	5938 (0.40)	11250 (0.76)
JARO_2	880	863 (0.98	842 (0.98)	842 (0.96)	842 (1)	841 (1)	828 (0.98)	402 (0.48)	339 (0.40)	360 (0.43)	631 (0.75)
JARO_31	3709	3559 (0.96)	3511 (0.99)	3511 (0.95)	3511 (1)	1093 (0.31)	3511 (1)	3510 (1)	1362 (0.39)	2888 (0.83)	3510 (0.99)
BARO_3	3716	3581 (0.96)	3530 (0.99)	3530 (0.95)	2629 (0.75)	1140 (0.32)	1869 (0.53)	1618 (0.46)	1363 (0.39)	1384 (0.39)	3512 (0.99)
BARO_10	566	430 (0.76)	422 (0.98)	422 (0.75)	337 (0.80)	137 (0.32)	248 (0.59)	191 (0.46)	156 (0.37)	422 (1)	199 (0.47)
HIAN_3	5402	5284 (0.98)	5221 (0.99)	5221 (0.97)	5221 (1)	1807 (0.34)	5218 (1)	2197 (0.42)	1848 (0.36)	5184 (0.99)	2291 (0.43)
HIAN_7	1199	1171 (0.98)	1142 (0.98)	1142 (0.95)	1142 (1)	363 (0.32)	597 (0.53)	519 (0.46)	423 (0.37)	474 (0.42)	500 (0.43)
SEMI_2	2230	2136 (0.98)	1969 (0.92)	1969 (0.88)	476 (0.24)	533 (0.27)	1969 (1)	847 (0.43)	664 (0.34)	592 (0.30)	1470 (0.75)
SEMI_24	4980	4397 (0.88)	3473 (0.79)	3473 (0.70)	3059 (0.88)	1204 (0.35)	3473 (1)	2239 (0.65)	2030 (0.58)	1891 (0.54)	990 (0.29)
BROYV_35-1	388	260 (0.67)	230 (0.88)	230 (0.59)	230 (1)	230 (1)	223 (0.97)	214 (0.93)	210 (0.91)	194 (0.84)	140 (0.61)
BROYV_40-1	518	103 (0.20)	86 (0.83)	86 (0.17)	67 (0.78)	27 (0.31)	-	57 (0.66)	55 (0.64)	36 (0.42)	86 (1)
BROYV_41-2	1491	1300 (0.87)	1245 (0.96)	1245 (0.83)	1245 (1)	1245 (1)	1222 (0.98)	1234 (0.99)	1232 (0.99)	513 (0.41)	1245 (1)
GRPI_9	2950	2849 (0.97)	2806 (0.98	2804 (0.95)	2207 (0.79)	918 (0.33)	1532 (0.55)	1354 (0.48)	1130 (0.40)	1162 (0.41)	1267 (0.45)
GRPI_40	5091	2961 (0.58)	2883 (0.97)	2883 (0.57)	2179 (0.76)	941 (0.33)	2846 (0.99)	2845 (0.99)	2239 (0.78)	1540 (0.53)	1669 (0.58)
BAGA_10	879	764 (0.87)	699 (0.91)	699 (0.80)	699 (1)	640 (0.92)	673 (0.96)	670 (0.96)	662 (0.95)	574 (0.82)	549 (0.79)
BAGA_28	2024	1624 (0.80)	1515 (0.93)	1515 (0.75)	1515 (1)	1507 (0.99)	1515 (1)	737 (0.48)	602 (0.40)	770 (0.51)	1515 (1)

5.3 Pipeline `ViroPhylo`

5.3.1 Introduction

Il existe chez un nombre important de polyomavirus humains, tels que les BKPyV, JCPyV et MCPyV, une phylogénie où l'on retrouve différents niveaux taxonomiques que l'on peut qualifier de groupes ou de sous-groupes en fonction de leur profondeur [199]. Il a été montré qu'une infection par un polyomavirus peut être plus ou moins symptomatique selon le génotype prédominant. Dans le cas d'une PVAN, *Nukuzuma et al.* ont montré que les infections par le génotype I des BKPyV sont davantage cytopathiques au sein des cellules épithéliales du rein que celles provoquées par le génotype IV [281]. Il est donc nécessaire de détecter par quel génotype un patient à risque est infecté, pour prévoir les risques différentiels d'une PVAN.

Il existe plusieurs méthodes de génotypage des virus, allant de la reconstruction phylogénétique sur un jeu de données de référence à la mise en place de clefs dichotomiques de génotypage sur des régions *hotspots* (point chaud en français, *i.e.* régions déterminantes des groupes taxonomiques). Chez les polyomavirus, c'est l'analyse de ces *hotspots* qui est fortement utilisé dans le contexte clinique.

La majeure partie des génotypages de BKPyV dans le contexte clinique sont effectués sur le gène *VP1*. Il a été établi par *Jin et al.* que la détermination des groupes de BKPyV pouvait être assurée par l'analyse d'une région de 327 nucléotides (appelée région de *Jin*) [282]. Cela dit, *Luo et al.* émettent des réserves sur l'utilisation de cette région pour le génotypage, et mettent en avant l'utilisation systématique du gène *LT* pour effectuer une telle tâche [200]. En effet, ils présentent au sein de leur article une reconstruction phylogénétique de 178 BKPyV à partir la région de Jin et trouvent de fortes incohérences au sein de la détermination des sous-groupes du BK IV. Toujours pour promouvoir l'utilisation du gène *LT* dans le génotypage des BKPyV, ces mêmes auteurs comparent les résolutions phylogénétiques obtenues à partir des gènes *LT* et *VP1* de 178 polyomavirus. Ils en déduisent que bien que les deux phylogénies soient similaires, celle construite avec le gène *VP1* possède une topologie moins robuste [200]. Bien que les hésitations soient de mise entre l'utilisation de *VP1* et de *LT* pour le génotypage des polyomavirus, le gène *VP1* reste toutefois favori. Récemment, une étude méthodologique de *Morel et al.* a mis en place un algorithme dichotomique de décision pour déterminer le génotype d'un BKPyV sur la base de l'analyse de la région BKTGR du gène *VP1*. Cette région est composée de seulement 100 nucléotides et selon ses auteurs, son analyse permettrait d'identifier avec une efficacité de 99 % les BKPyV. L'analyse de la région BKTGR constitue donc à ce jour l'une des méthodes les plus poussées pour génotyper les BKPyV [283]. Mais pour aller plus loin dans les solutions de génotypage de BKPyV, il est nécessaire de proposer une solution universelle accessible, fonctionnant sur n'importe quel jeu de données de BKPyV et dont l'efficacité serait encore plus poussée.

Nous présentons au sein de cette section l'outil `ViroPhylo`, un pipeline utilisant les programmes `BLAST` [227], `MAFFT` [260] et `RAxML` [234] pour permettre un génotypage rapide et efficace des polyomavirus humains avec un approfondissement de l'analyse pour les souches BK et JC.

5.3.2 Fonctionnement de ViroPhylo

L'objectif principal de ViroPhylo est d'incorporer un jeu de données de séquences de polyomavirus au sein d'une référence qui comprend un alignement et un arbre phylogénétique associé. Il s'agit d'un pipeline *standalone* (*i.e* pouvant fonctionner sans l'aide d'un autre programme) qui sera prochainement disponible sur le site http://virophylo.ird.fr. L'architecture de ViroPhylo est donnée en Figure 5.2.

Dans un premier temps, ViroPhylo vérifie et modifie les séquences entrées par l'utilisateur pour qu'elles forment un jeu de données exploitable par le pipeline. Une première étape utilisant la fonction blastn de l'outil BLAST compare les séquences de l'utilisateur à des gènes de référence et permet d'y retrouver la présence d'ADN d'intérêt [227]. Dans le cas où les séquences sont bien issues de polyomavirus, elles font l'objet, si besoin, d'un *reverse-complement* de manière à ce que leur matériel génétique soit dans le sens 5'-3' du brin transcrit. Ensuite, dans le cas où le matériel génétique donné en entrée est constitué de génomes, chaque séquence fait l'objet d'une analyse par le sous-programme ROTATOR. Celui-ci permet la rotation des génomes de polyomavirus pour qu'ils commencent tous au premier codon de *VP2* : la séquence est répétée deux fois et les trois réplicas sont concaténés. De cette manière, chaque séquence nouvellement construite possède un exemplaire de la séquence d'origine enclavée entre deux gènes *VP2*. Un nouveau blastn permet de détecter les différentes occurrences de *VP2* et leur position. En sélectionnant uniquement la séquence située entre le premier codon de la deuxième occurrence de *VP2* et la troisième, nous récupérons un génome complet de polyomavirus débutant exactement au premier nucléotide du gène *VP2*. La raison pour laquelle nous choisissons le gène *VP2* comme point de départ de tous les génomes est que ce gène est non seulement partagé par tous les polyomavirus, mais aussi que l'organisation des génomes, sauf exception, peut parfaitement s'articuler autour de cette conformation. Les séquences nettoyées et préparées sont comparées une nouvelle fois à des génomes de référence de polyomavirus pour vérifier que la majorité du contenu génétique se rapproche des polyomavirus connus. Une fois cette étape vérifiée, les séquences d'entrée sont alignées à un jeu de données de référence déjà aligné correspondant au niveau de génotypage souhaité par l'utilisateur. En effet, l'utilisateur peut choisir d'effectuer son analyse ViroPhylo sur un alignement de référence correspondant au génome complet, à une concaténation de gènes (précoces, tardifs ou de tous les gènes) ou encore à un gène en particulier sur les BKPyV, JCPyV, les polyomavirus humains ou sur tous les polyomavirus. Cette étape d'alignement est réalisée grâce à l'option « add » de MAFFT [260]. L'alignement obtenu est ensuite inséré au sein d'une analyse phylogénétique réalisée par l 'outil RAxML (v8.1), et ce en vue d'ajouter les séquences requêtes à une phylogénie que nous avons construit au préalable et qui correspond à l'alignement de référence auquel les séquences ont été ajoutées (voir chapitre précédent). Pour cela, nous utilisons la fonction « epa » (pour *evolutionary placement algorithm*) de RAxML, qui permet de situer les séquences requêtes sur une phylogénie de référence associée à l'alignement de référence [234]. Les résultats du placement sont donnés à l'utilisateur avec une probabilité associée. À la lecture de la table des probabilités associées, l'utilisateur peut connaître, avec une certaine robustesse, dans quel clade s'insèrent les séquences requêtes. Il est à noter que les génomes de polyomavirus sont peu sujets à de la recombinaison, ce qui nous permet d'assurer à un certain degré le génotypage.

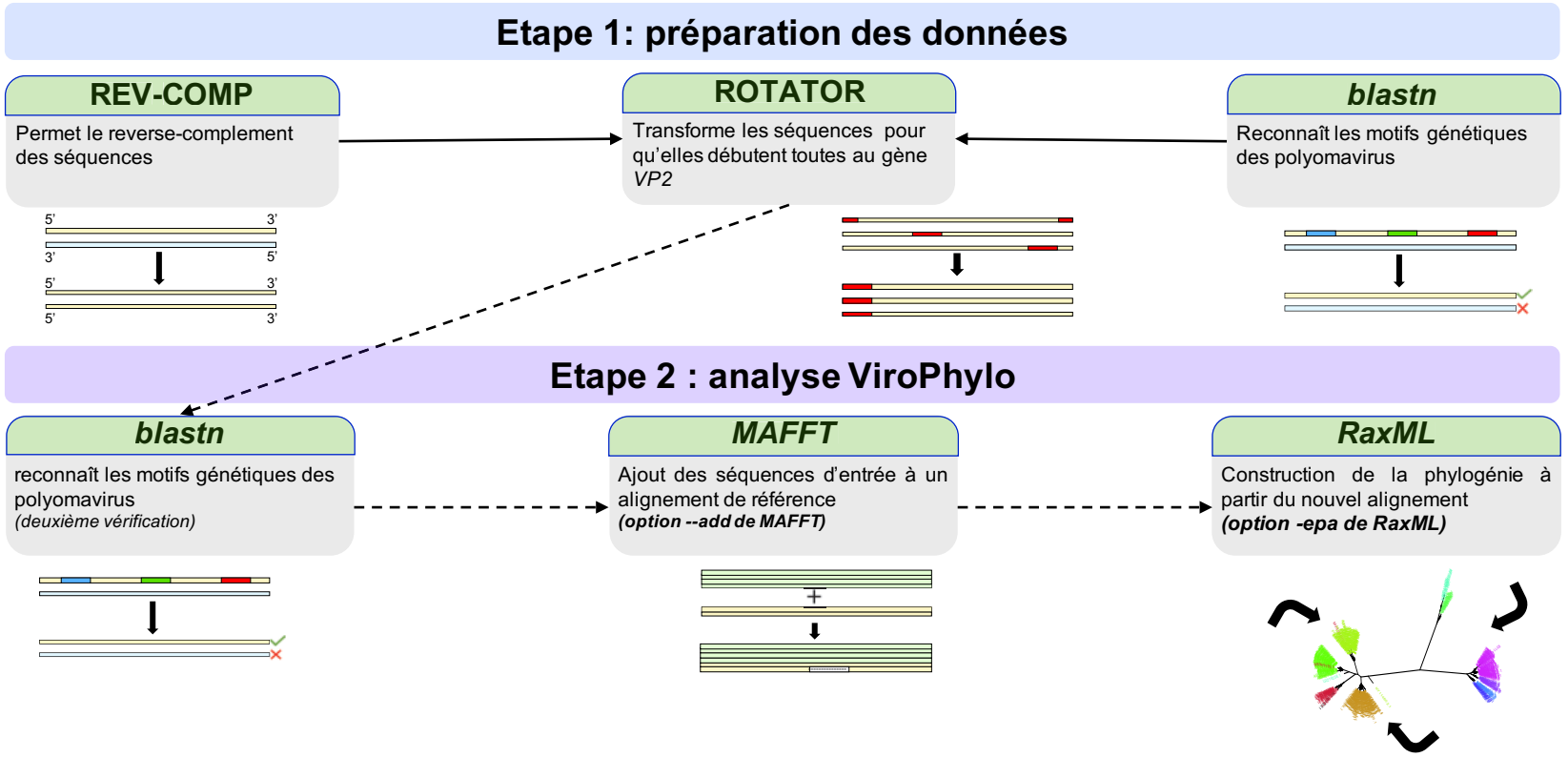

Figure 5.2 – **Pipeline de l'outil** `ViroPhylo`**.** Le pipeline `ViroPhylo` peut traiter des données sur un génome complet ou sur une partie de celui-ci. Une fois les données d'entrée insérées, une première étape prépare les données de l'utilisateur comme suit : i) analyse des données sur des gènes de référence à l'aide de la fonction `blastn` de l'outil `BLAST` [227]. Cette étape a pour objectif de vérifier que les données d'entrée soient bien relatives à des polyomavirus ii) Action de *reverse-complement* sur les données n'ayant pas une direction 5'-3' des gènes tardifs iii) rotation des données (s'il s'agit de génomes) pour que celles-ci débutent au premier codon du gène *VP2*. Une fois les données préparées pour l'analyse, une nouvelle étape de vérification par *blast* sur référence permet de vérifier si les données sont des génomes et de confirmer, encore une fois, que les données analysés appartiennent aux polyomavirus et qu'elles soient dans la direction 5'-3' des gènes tardifs. Par la suite, ces données sont rajoutées à un alignement de référence, et ce par la fonction « -add » de l'outil MAFFT [260]. Ce nouvel alignement est analysé par l'outil `RAxML` pour placer les nouvelles données sur un arbre phylogénétique de référence grâce à la fonction « epa » de ce même outil [234].

5.3.3 Avantages du pipeline `ViroPhylo`

En plus de proposer une solution simple et universelle pour génotyper les polyomavirus humains, `ViroPhylo` possède une exactitude jusqu'ici inégalée. En prenant pour exemple comparatif l'algorithme dichotomique de *Morel et al.*, qui génotype faussement cinq individus sur les 493 analysés, `ViroPhylo` permet un génotypage sans erreur sur le même jeu de données (Figure 5.3). De plus, de par l'utilisation d'une phylogénie de référence comme base pour le génotypage, `ViroPhylo` pourrait à l'avenir permettre de déterminer l'existence de nouvelles souches de polyomavirus humains.

5.4 Modélisation de l'évolution intra-hôte du BKPyV dans le cadre d'une PVAN

L'objectif préliminaire de ce projet entrait dans la détermination et l'analyse de l'évolution des génomes de BKPyV chez un patient ayant reçu une greffe de reins et ayant été immunosupprimé pour prévenir un rejet par le système immunitaire. Au sein de cette partie du projet, nous avons tenté de mettre en place une amélioration du modèle d'évolution intra-hôte de *Funk et al.* [1]. Malheureusement, il n'a pas été possible de reproduire le modèle de base de *Funk et al.*. Après plusieurs tentatives infructueuses de contact avec le créateur de ce modèle, mais aussi l'absence de données virales pour notre projet, nous l'avons temporairement mis de côté pour nous focaliser sur d'autres aspects de ce doctorat.

5.4.1 Introduction

L'évolution et la diversité intra-hôte des BKPyV lors de la rémission d'une greffe de reins demeure un aspect primordial pour notre compréhension des PVAN et échecs de greffe de reins [1]. Plusieurs auteurs ont tour à tour tenté de modéliser la dynamique d'infection chez un patient immunodéprimé, avec une focalisation sur les populations virales [1].

En 2008, le modèle de *Funk et al.* a été développé dans l'objectif de déterminer la cinétique de l'évolution de la virémie et de la virurie lors d'une infection par les BKPyV, et ce à la suite d'une greffe de rein [1]. Il s'agit d'un modèle à deux compartiments où le premier est défini par les cellules du rein et le second par celles du système urinaire. Chaque cellule d'un compartiment peut être saine ou infectée par le BKPyV. La dynamique de transmission d'une cellule à une autre est définie par des paramètres intra-compartimentaux et inter-compartimentaux. Dans le dernier cas, la transmission est représentée par l'efflux (allant du rein vers le système urinaire) ou le reflux (allant du système urinaire au rein) des BKPyV. Il est à noter que les cellules saines peuvent se multiplier et mourir sans le concours des virus et que la mort des cellules infectées conduit automatiquement à un *burst* infectieux associé à la libération des virions qui sont alors capables d'infecter une autre cellule saine. Ce modèle, de prime abord simple, a été testé sur des données réelles et sous plusieurs conditions initiales. À la suite d'une analyse longitudinale de la virémie et de la virurie au sein de patients infectés par le BKPyV, *Funk et al.* confirmèrent l'exactitude de leur modèle et profitèrent de sa fiabilité pour confirmer qu'un traitement par immunomodulation est nécessaire dès lors que la virémie atteint un seuil de 10^4 copies/mL d'ADN viral, et que celui-

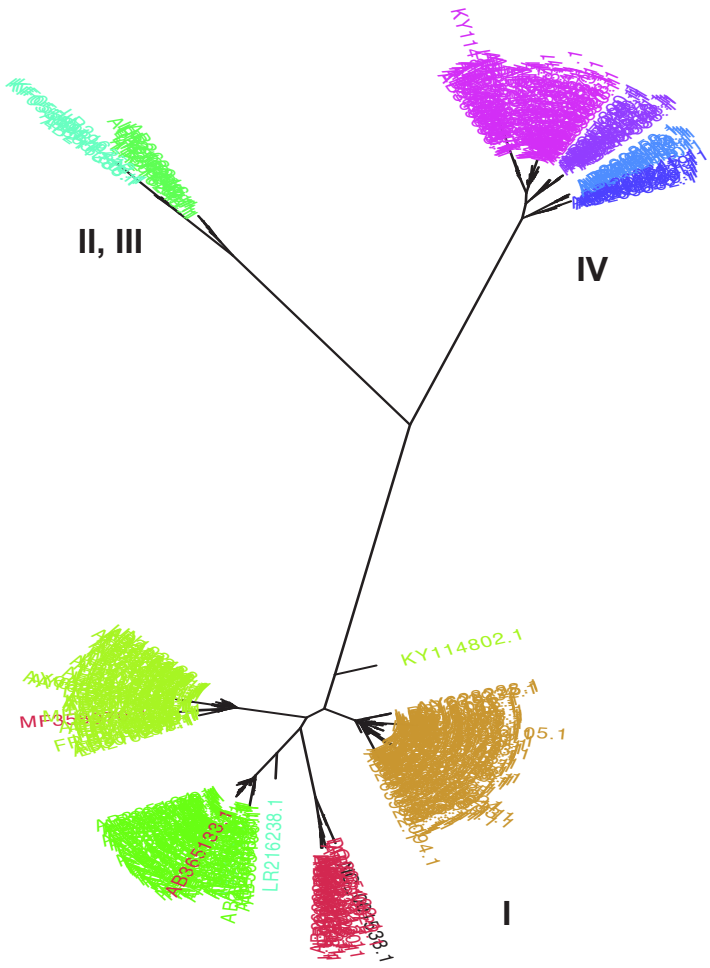

Figure 5.3 – **Phylogénie des BKPyV de l'outil** `ViroPhylo`**.** Cette phylogénie a été construite à partir des CDS complets de 493 génomes BKPyV (données issues de la banque de données NCBI [227]). L'inférence phylogénétique a été faite sur un modèle de Maximum de Vraisemblance GTRGAMMA avec un bootstrap de 1000 et ce à l'aide du prgramme `RAxML` [234]. Les couleurs données au sein de ce graphe représentent le génotypage proposé par *Morel et al.*. Certaines séquences ne rejoignent pas les clades possédant une majorité de séquences partageant la même couleur et peuvent donc être considérés comme mal génotypées. Au final, 5 individus sur 493 sont faussement génotypés par l'algorithme de *Morel et al.*, ce qui représente environ 1 % des génomes ici étudiés.

ci doit faire baisser 80 % de la virémie pour être efficace [1]. Parmi les scénarios possibles quant à l'origine de la réactivation du virus, ces mêmes auteurs proposèrent une hypothèse *kidney-urinary tract-kidney*, où le virus se réplique de manière précoce et peu invasive dans le rein greffé, prolifère dans le compartiment urinaire et provoque la PVAN par reflux des virions dans le rein. Ce scénario suggèrerait alors que le BKPyV réactivé proviendraient du donneur, soupçon confirmé par la suite par d'autres auteurs [1, 202, 220].

Les symboles utilisés dans le modèle de *Funk et al.* sont décrits au sein du Tableau 5.4.1. Ce modèle est composé de deux compartiments comprenant six variables :

$$\frac{dE}{dt} = p_1 E \left(1 - \frac{E}{E_{\max}}\right) - d_1 E - i_1 I E - r V_u E \tag{5.1}$$

Cette équation représente la dynamique des cellules saines E du rein, où p_1 représente le taux de croissance, E le nombre de cellules saines au temps t, $E_{\max}$ le nombre maximum de cellules saines dans le compartiment, d_1 le taux mortalité de cellules saines non infectées, i_1 le taux d'infection des cellules saines, I le nombre de cellules infectées, r le taux de reflux des virions et V_u le nombre de virions au sein du tractus urinaire. On distingue plusieurs voies d'infection. Une par les virions du système rénal (décrite par $i_1 I E$) et l'autre par les virions du système urinaire (décrite par $r V_u E$).

$$\frac{dI}{dt} = i_1 I E + r V_u E - u_1 I \tag{5.2}$$

Cette équation représente la dynamique des cellules I infectées du rein, où u_1 représente la mortalité des cellules infectées.

$$\frac{dV_p}{dt} = i s(t) b_1 u_1 I - c_1 V_p \tag{5.3}$$

Cette équation représente la dynamique des virions V_p du rein, où $is(t)$ représente le système immunitaire de l'hôte, b_1 le nombre de virions produits par une cellule infectée et c_1 la mortalité des virions.

$$\frac{dU}{dt} = p_2 U \left(1 - \frac{U}{U_{\max}}\right) - d_2 U - i_2 Y U - e V_p U \tag{5.4}$$

Cette équation représente la dynamique des cellules saines U du compartiment urinaire, où p_2 représente le taux de croissance, U le nombre de cellules saines au temps t, $U_{\max}$ le nombre maximum de cellules saines dans le compartiment, d_2 le taux mortalité de cellules saines non infectées, i_2 le taux d'infection des cellules saines, Y le nombre de cellules infectées et e le taux d'efflux des virions. On distingue plusieurs voies d'infection. Une par les virions du système urinaire (décrite par $i_2 U Y$) et l'autre par les virions du système rénal (décrite par $e V_P U$).

TABLE 5.2 – **Symboles utilisés au sein du modèle de *Funk et al.*.** Sont décrits au sein de ce tableau le symbole, sa définition, et les valeurs associées [1].

Symbole	**Définition**	**Valeurs estimées**
E	Cellules saines du rein	conditions initiales
$E_{\max}$	Nombre maximal de cellules saines du rein	1.10^4
I	Cellules infectées du rein	$I_0 = 0$ ou 1.10^{-6}
U	Cellules saines du système urinaire	Conditions initiales
$U_{\max}$	Nombre maximal de cellules saines du système urinaire	1.10^7
Y	Cellules infectées du système urinaire	$Y_0 = 0$ ou 1.10^{-6}
p_1	Taux de croissance des cellules du rein	0.6
p_2	Taux de croissance des cellules du système urinaire	$1.2 - 1.5$
d_1	Taux de mortalité des cellules saines du rein	0.003
d_2	Taux de mortalité des cellules saines du système urinaire	0.006
u_1	Taux de mortalité des cellules infectées du rein	0.33
u_2	Taux de mortalité des cellules infectées du système urinaire	0.33
i_1	Taux d'infection des cellules saines du rein	2.10^{-7}
i_2	Taux d'infection des cellules saines du système urinaire	1.10^{-7}
V_p	Virions du rein	Conditions initiales
V_u	Virions du système urinaire	Conditions initiales
$is(t)$	Valeur d'immunosuppression	$1.10^{-4} - 1$
b_1	Nombre de virions relâchés après *burst* des cellules du rein	6000
b_2	Nombre de virions relâchés après *burst* des cellules du système urinaire	6000
c_1	Mortalité des virions du rein	10
c_2	Mortalité des virions du système urinaire	1
e	Efflux des virions (rein -> système urinaire)	1.10^{-9}
r	Reflux des virions (système urinaire -> rein)	1.10^{-7}
s	Efflux des virions sans infection directe des cellules U	$1 - 1000$

$$\frac{dY}{dt} = i_2 YU + eV_pU - u_2Y \tag{5.5}$$

Cette équation représente la dynamique des cellules Y infectées du système urinaire, où u_2 représente la mortalité des cellules infectées.

$$\frac{dV_u}{dt} = is(t)b_2u_2Y + sV_p - c_2V_u \tag{5.6}$$

Cette équation représente la dynamique des virions V_u du système urinaire, où b_2 représente le nombre de virions produits par une cellule infectée, s le taux d'efflux des virions du rein (sans qu'il n'infectent une cellule du système urinaire) et c_2 la mortalité des virions.

Une représentation schématique de ce modèle est donnée en Figure 5.4.

5.4.2 Intégration et améliorations possibles du modèle

Le modèle de *Funk et al.* ne prend pas en compte l'évolution et la diversité virale. L'ajout d'une diversité de populations de BKPyV dans le modèle permettrait de représenter différentes populations virales avec des valeurs de paramètres qui leur sont propres. L'inspiration de cette dernière étape dans le modèle vient d'une catégorisation de variants viraux décrite au sein de la revue de *Mazalrey et al.* [171]. Celle-ci rassemble les BK en types de souches mutantes ayant des différences dans leurs paramètres infectieux, notamment à cause de mutations dans leur région NCCR ou VP1. En effet, il a été montré par [284] que l'apparition de réarrangements des NCCR au sein de la population virale des BKPyV au cours d'une infection chez un patient immunosupprimé provoquait une augmentation de la réplication virale [171, 284, 285]. Ce même auteur a aussi remarqué que la prolifération des BKPyV était marquée par l'apparition de mutants sur la région *VP1*, ce qui permettrait un mécanisme d'échappement de l'immunité de l'hôte lors de sa réactivation pour combattre l'infection, et que les BKPyV exploiteraient les « lacunes » du système immunitaire pour permettre une installation durable [284, 285]. Il a été estimé que les mutations de *VP1* sont principalement dues à l'action de la désaminase de l'hôte APOBEC3B [284, 285]. Dans les deux cas, cette soudaine diversité n'apparaîtrait que lorsque l'immunité de l'hôte est réduite, et ne se retrouvent donc pas à l'état naturel [284, 285] [286]. Il est par ailleurs intéressant de noter que le taux de mutation observé chez les BKPyV semble augmenter chez les patients receveurs d'une greffe de reins par rapport à un hôte immunocompétent [286]. L'amélioration proposée pour notre modèle se focalise sur les réarrangements de la NCCR et leurs implications sur la modulation de la réplication virale. Il existerait alors trois catégories de mutants que l'on ne retrouverait que chez les patients immunosupprimés ; les deux premiers verraient l'expression de leurs gènes précoces ou tardifs augmenter alors que le troisième possède une l'expression accrue de tous ces gènes. Les souches décrites ci-dessus ne seraient alors pas sélectionnées dans le cadre d'une infection classique [171]. L'ajout de ces différents mutants au modèle de *Funk et al.* implique la prise en compte de différentes cellules infectées pour chaque compartiment. Comme chaque cellule peut être infectée par la souche *wild-type* ou l'un des

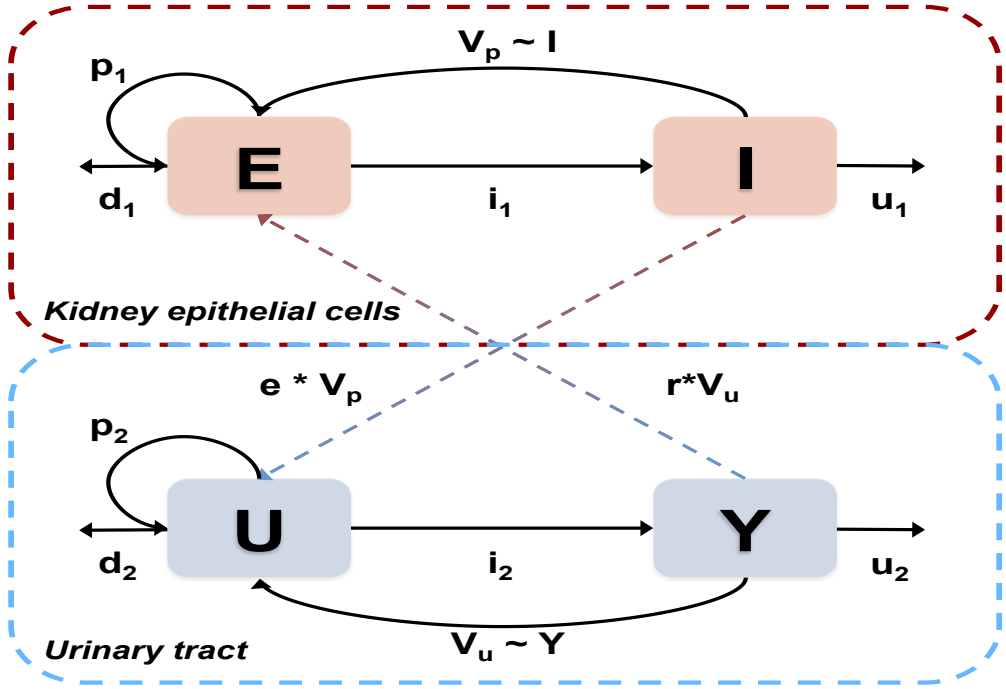

Figure 5.4 – **Modèle à deux compartiments de la dynamique virale des BKPyV.** Les deux compartiments représentent les cellules épithéliales du rein (encadré rouge) et du tractus urinaire (encadré bleu). Les sous compartiments E et I représentent les cellules saines et infectées du rein, et les cellules U et Y les cellules saines et infectées du tractus urinaire. Les flèches pleines représentent les paramètres de dynamique propre à chaque compartiment, et les flèches en pointillées les dynamiques d'efflux (rouge vers bleu) et de reflux (bleu vers rouge) des BKPyV. Un détail de chaque paramètre est indiqué dans le texte et au sein du Tableau 5.4.1. Cette figure est inspirée de l'article de *Funk et al.* [1]

mutants, il existe alors quatre différentes cellules infectées, chacune avec ses propres valeurs de paramètres associés.

À titre d'exemple, si l'on prend les cellules infectées I du rein, on aura, pour un des mutants, l'équation suivante :

$$\frac{dI'}{dt} = i_{1'}IE + rV_uE - u_{1'}I' \tag{5.7}$$

Où tous les paramètres I', $i_{1'}$, $u_{1'}$ posséderaient les valeurs spécifiques au mutant, probablement définies par un taux d'infection des cellules saines et de mortalité des cellules infectées supérieurs au *wild-type*.

La souche virale $V_{p'}$ associée serait quant à elle définie par l'équation suivante :

$$\frac{dV_{p'}}{dt} = is(t)b_{1'}u_{1'}I - c_{1'}V_{p'} \tag{5.8}$$

Où, encore une fois, les paramètres $b_{1'}$, $u_{1'}$ et $c_{1'}$ seraient définis par les spécificités de la souche virale observée.

Bien entendu, la dynamique des cellules saines serait elle aussi modifiée :

$$\frac{dE}{dt} = p_1 E\,(1 - E/E_{\max}) - d_1 E - i_1 IE - i_{1'} I'E - rV_u E - rV_{u'} E \tag{5.9}$$

Où $rV_{u'}E$ représente le reflux des virions de la souche virale $V_{u'}$.

Ainsi, nous aurions un modèle complexifié par l'apparition de trois nouveaux sous-compartiments dans le rein, et de trois autres dans le système urinaire, modifiant par la même occasion les populations virales des deux compartiments. Par la simulation d'une diversité, nous espérons affiner le modèle proposé par *Funk et al.* en proposant un approfondissement de la dynamique virale au cours d'une infection [1]. Une fois ce modèle mis en place, celui-ci sera couplé aux données génétiques, de virurie et de virémie chez les patients receveurs d'une greffe de rein de l'ANR BK-NAB.

5.5 Discussion

Pour mieux comprendre l'évolution de la réactivation post-greffe des BKPyV au sein d'un patient, il est nécessaire de se pencher sur les mécaniques définissant l'évolution et la diversité intra-patient du virus. D'après plusieurs études, celle-ci est marquée par l'installation de souches mutantes au sein des populations virales qui ne sont pas détectées à l'état naturel au sein de personnes saines [171, 284–286]. Ces souches mutantes permettraient non seulement de mettre en place une diversité permettant un évitement du système immunitaire de l'hôte, mais aussi un accroissement de la prolifération virale [171, 284–286]. Au sein de ce chapitre, nous proposons un panel d'outils et de modèles pour permettre un traitement aisé de données associées au BKPyV. Ceux-ci visent principalement à déterminer les spécificités génétiques des BKPyV contenus au sein d'un échantillon, que ce soit pour les génotyper (avec `ViroPhylo`) ou pour effectuer un état de la diversité des BKPyV à l'échelle génomique et des gènes (avec `GenoPolys`). Les données génétiques des BKPyV, en relation avec les données personnelles d'un patient ainsi que de virémie et de virurie permettrait de mieux comprendre les profils des BKPyV provoquant une PVAN. l'ANR BK-NAB veut poursuivre ces objectifs de manière systématique par l'analyse de données longitudinales sur des patients receveurs d'une greffe de reins. Aujourd'hui, `GenoPolys` et `ViroPhylo` sont déjà utilisés par le CHU de Nantes et feront l'objet d'une publication méthodologique.

Chapitre Six

6 Conclusion et Perspectives

Le projet initial de ce doctorat était de déterminer les dynamiques virales de patients atteints d'une PRR (due aux papillomavirus humains HPV6 et HPV11) ou d'une PVAN (après greffe de reins ; maladie causée par le BKPyV). Pour des raisons indépendantes de notre volonté, nous n'avons pas pu obtenir de données concernant la dynamique virale de patients atteints de la PRR et les données relatives à la PVAN sont toujours en cours d'obtention. Pour préparer l'analyse des données issues de l'ANR BK-NAB, nous avons conçu deux pipelines d'analyse des données génétiques de polyomavirus humains et les avons confrontés à des données préliminaires issues des collectes longitudinales sur les patients receveurs d'une greffe de rein du CHU de Nantes. Ces deux pipelines, déjà utilisés par la plate-forme du CHU de Nantes, permettront à l'avenir un traitement rigoureux de données des polyomavirus BK. À l'avenir, une étude systématique de la diversité génétique sera effectuée le long de la rémission des patients. Celle-ci se focalisera sur l'évolution de la région NCCR (substitutions et indels au niveau des régions O, P, Q, R et S) et du gène *VP1* (principalement sur les substitutions). L'objectif sera d'observer si, au cours d'une évolution du diagnostic de PVAN, les profils génétiques des polyomavirus BK sont modifiés. Par ailleurs, cette collecte de données longitudinales permettra d'observer la dynamique de ces virus, et nous permettra de mieux comprendre et prévenir une infection pouvant provoquer une PVAN.

De par les mécanismes qu'il peut impliquer sur les étapes transcriptionnelles et traductionnelles de la machinerie cellulaire, le CUB est un élément primordial à prendre en compte pour améliorer notre compréhension des processus liés à l'expression des gènes. Aujourd'hui, suffisamment de connaissances ont été accumulées pour démontrer que le CUB est à l'origine de mécanismes de régulation de l'expression des gènes, mais qu'il peut aussi être influencé par des biais mutationnels. La régulation de l'expression des gènes par le CUB peut, par exemple, s'opérer *via* des modifications sur la structure secondaire des ARNm, de leur stabilité et de leur épissage [64–67, 69–71], ou encore sur le rythme et de la fidélité de l'étape d'élongation de la traduction [2, 31, 37–39, 55, 72, 77] et, pour finir, sur le repliement des protéines [86–89]. Il peut être influencé par des mécanismes tels que le gBGC [43, 107], le TAMB [100, 101], ou tout simplement le biais mutationnel universel vers les nucléotides AT [92, 96, 97]. La sélection et le biais mutationnel définissent à tous deux une mosaïque de mécanismes façonnant la composition nucléotidique des gènes. De telles découvertes, sous un certain angle, remettent en cause la

sémantique même du terme « codons synonymes », et sont en grande partie dues aux avancées technologiques de ces vingt dernières années. La mise en place de méthodes « haut-débit » en génomique, en transcriptomique et en protéomique a permis l'analyse des particularités cellulaires de nos organismes, et a participé à la confirmation de certaines hypothèses quant au rôle du CUB dans l'expression des gènes. À l'exception des Vertébrés, mais surtout des mammifères, les mécanismes qui définissent le CUB sont largement expliqués par une combinaison des deux origines mutationnelles et traductionnelles, ce qui semble souligner une véritable universalité des mécanismes du CUB au sein du vivant.

Au cours de ce doctorat, nous avons développé COUSIN, un indice de mesure du biais d'usage des codons novateur, et proposant des résultats avec une signification biologique immédiate. COUSIN est un indice avant tout comparatif (il compare le CUB d'une requête à une référence) mais son plus grand avantage réside dans le fait que les résultats comparatifs sont normalisés autour d'une hypothèse nulle H_0 décrivant un usage équiprobable des codons synonymes [94]. Grâce à cette particularité, COUSIN propose une plage de scores se rapprochant d'une véritable mesure du CUB, *a contrario* de l'indice CAI qui mesure l'« optimalité » d'une séquence par rapport à une référence [47]. De plus, cette normalisation permet une uniformité dans les résultats obtenus, principalement sur un point de vue qualitatif. Contrairement à l'indice CAI, les valeurs clefs de COUSIN (0 pour un CUB similaire à l'H_0, 1 pour un CUB similaire à la référence) et la position qualitative du score de la requête par rapport à ces valeurs reste inchangée quelle que soit la nature de la référence. De ce fait, COUSIN permet non seulement de comparer le CUB d'une séquence par rapport à une référence, mais invite à une comparaison, par exemple, inter-espèces de gènes orthologues.

Si les données nouvellement acquises par les technologies « haut-débit » ont permis de confirmer certains aspects du rôle du CUB dans l'expression des gènes chez certains procaryotes [31, 54], eucaryotes unicellulaires [38] et invertébrés [39, 55, 58], le débat reste toujours ouvert quand aux implications de la sélection traductionnelle chez les Vertébrés [36, 43]. Pour être plus précis, ce sont plutôt les hypothèses de sélection positive des codons synonymes pour améliorer le processus de traduction, et ce en accord avec les populations d'ARNt de la cellule, qui sont ici le sujet d'un débat [36, 43, 287, 288]. Mais la présence écrasante du biais mutationnel au sein de ces organismes pourrait peut être bien masquer des mécanismes de sélection actant sur le CUB des gènes [36, 43]. Peu d'études s'attardent sur le CUB avec une notion de nivellement spatio-temporel sur les données transcriptomiques et protéomiques, où l'expression des gènes et l'abondance des ARNt des différents tissus et étapes du cycle cellulaire sont pris en compte dans les analyses. Dans bien des cas, les études se basent sur une estimation du CUB avec comme référence le CUB du génome complet [32]. Pourtant, si sélection traductionnelle il y a chez les Vertébrés, le CUB devrait avant tout être observé dans les données transcriptomiques et protéomiques. Les études de *Chevallier et Garel*, portant sur l'expression temporelle des gènes de *Bombyx mori* et sur l'influence du CUB et des populations d'ARNt sur celle-ci avaient déjà permis de considérer une telle approche pour la mesure du CUB et sur sa possible relation avec l'expression des gènes [57]. *Chevallier et Garel* proposèrent une étude spatiale (sur les différents tissus) et temporelle (lors du développement larvaire) de l'organisme, et réussirent à établir un

lien entre CUB, expression des gènes et populations en ARNt. Peu de temps après, *Hastings et Emerson* proposèrent une étude sur les gènes spécifiques des tissus du foie et du muscle squelettique communs à plusieurs organismes aviaires et mammifères, sans pour autant observer des particularités dans le CUB observé entre les des deux tissus [59]. Techniquement, ces deux études ouvrèrent les portes d'une analyse spatio-temporelle du CUB chez les Invertébrés et les Vertébrés, et permirent toutes deux de considérer la relation entre CUB, ARNt et expression des gènes sous un angle nouveau.

C'est sur cette philosophie que nous nous sommes basés pour l'analyse des gènes paralogues *PTBP*. Notre objectif était de déterminer les spécificités nucléotidiques de séquences homologues dans leur fonction mais pas dans leur CUB ni dans leur expression, et de tenter de détecter des signatures d'une sélection des codons synonymes au sein des différents gènes paralogues au cours de leur évolution. Pour résumer, nous visions à observer si l'effet du contexte génomique, généralement explicatif du CUB des gènes de Vertébrés, pouvait être différent selon le paralogue observé, et si cette différence pouvait être due à une sélection traductionnelle actant sur l'expression des paralogues au sein des tissus de mammifères [43]. Il a été remarqué que le CUB des gènes *PTBP1*, qui sont exprimés au sein de la majorité des tissus, était corrélé à la composition nucléotidique des introns et des régions flanquantes, soulignant alors un effet primordial des biais mutationnels sur le CUB observé [43]. En contre-partie, on observe une faible corrélation entre contenu en GC des régions non-codantes et contenu en GC3 du paralogue *PTBP2*. Il serait intéressant de poursuivre ce type d'analyse sur deux axes distincts. Tout d'abord, une analyse systématique des paralogues partageant une même fonction et une identité à l'échelle des protéines similaire à celle des *PTBP* permettrait de vérifier si ce genre de tendance est observé de manière constante chez les gènes paralogues des Vertébrés. D'un autre côté, une série d'études sur le même modèle que *Robinson et al.*, à l'échelle des organismes Vertébrés non-mammifères (où le CUB des trois paralogues est certes différent mais de manière discrète) et mammifères (ou le CUB des paralogues est exacerbé vers de l'enrichissement en GC ou en AT) permettrait de confirmer le rôle du CUB sur l'expression tissu-spécifique des gènes paralogues *PTBP* : si l'expression n'est pas modulée en fonction du CUB chez les Vertébrés non-mammifères, mais l'est chez les mammifères, cela pourrait bel et bien signifier un mécanisme de sous-fonctionnalisation des gènes par le CUB. Mais, bien que l'étude de gènes paralogues constitue un parfait terreau pour l'analyse du CUB et des possibles sélections traductionnelles qui le façonnent, il est toutefois nécessaire de pousser l'analyse à une échelle plus généraliste.

Les apports des études récentes menées par *Whittle et Extavour* et *Whittle et al.* pourraient définir un plan de recherche sur le CUB des Vertébrés en fonction du tissu étudié [39, 58]. Ces études visent à pondérer le CUB des tissus d'un organisme par l'expression de leurs gènes, et proposent donc une analyse spatiale du rôle du CUB tout en détectant non pas les codons synonymes « optimisés » au sein des gènes fortement exprimés, mais plutôt ceux dont la fréquence augmente le plus entre les gènes faiblement et fortement exprimés. Pour cela, il est nécessaire d'obtenir les données génomiques de l'organisme, mais aussi les données transcriptomiques et protéomiques de chacun des tissus et, pour aller plus loin, sous différentes contraintes et temporalités du cycle cellulaire. En effet, il a été démontré chez *H. sapiens*, mais aussi chez d'autres organismes, que la quantité et les populations des différents ARNt sont modifiées au cours du

cycle cellulaire ou sous certains stress [75, 76]. Même s'il ne s'agit dans le cas de l'étude portant sur *H. sapiens* que d'une modification globale de l'abondance des ARNt, et non pas de leurs proportions respectives, l'étude proposée par *Frenkel-Morgenstern et al.* indique clairement une baisse de l'expression des gènes utilisant des codons ne pouvant être associés qu'à des ARNt *near-cognate* [76]. De ce fait, les études précédentes qui se focalisaient uniquement sur l'aspect génomique du CUB, ou du moins en grande partie, se sont peut être fourvoyées en considérant le CUB comme un phénomène spatialement et temporellement uniforme. La question se pose de savoir s'il faut considérer le rôle du CUB dans les fluctuations temporelles, spatiales et conditionnelles, si elles existent, comme une preuve de l'existence d'une sélection traductionnelle, plutôt que d'un déséquilibre à l'échelle cellulaire pour lequel le CUB soudainement « optimisé » de certains gènes ne serait qu'une conséquence fortuite.

Pendant le déroulement de ce doctorat, nous avons visé au cours d'un projet annexe à analyser le CUB de 32 tissus humains (ces tissus sont présentés au sein du Tableau 6.2) selon les approches développées par *Whittle et Extavour* et *Whittle et al.* [39, 58]. Pour cela, nous avons estimé le CUB des gènes de chaque tissu en fonction de données RNA-seq issues d'une étude de *Wang et al.* [289]. Pour être plus précis, nous représentons les codons d'un gène autant de fois qu'il est représenté au sein des données RNA-seq (*i.e.*, leur valeur TPM pour *Transcripts Per Million*). Pour différencier les gènes au sein de chaque tissu, nous avons proposé une classification selon leur niveau d'expression : nous avons créé les catégories « Top » (gènes les plus exprimés parmi tous les tissus), « Bottom » (les moins exprimés parmi tous les tissus), Q1, Q3 et Q1_Q3 (gènes les 25 % moins exprimés par tissu, les 25 % les plus exprimés et le reste). Nous avons effectué une ACP sur le CUB des gènes des 32 tissus organisés selon cette classification (Figure 6.1). De manière intéressante, les deux premiers axes de l'ACP séparent le CUB des groupes de gènes selon leur expression au sein des différents tissus, et ce avec une puissance explicative de la variance d'environ 85 %. Le premier axe sépare les groupes de gènes selon leur contenu en GC, alors que le deuxième les sépare sans motif particulier. Il est à noter que bien que le premier axe permette l'isolation de certains groupes, comme les gènes des catégories « Top » et « Q1_Q3 » , c'est principalement le deuxième axe qui sépare avec précision les groupes de gènes selon leur niveau d'expression. Ainsi, les groupes « Top » , « Q3 » , et « Q1_Q3 » forment des groupes distincts dont le contenu en GC est légèrement variable, alors que les groupes de gènes faiblement exprimés « Bottom » et « Q1 » se confondent et possèdent un CUB différentiel selon le tissu concerné. À l'instar de l'étude sur les gènes *PTBP*, ces résultats préliminaires séparent encore une fois le codon UUG des autres codons riches en GC.

Sur ces mêmes données, nous avons effectué une brève analyse pour démarquer les codons synonymes ayant le changement le plus conséquent de fréquence entre les gènes fortement exprimés et les autres. Ces résultats sont présentés au sein de la Figure 6.2, et démontrent l'existence d'une différence dans les codons « optimisés » (comme l'entendent *Whittle et Extavour* et *Whittle et al.*) qui force à se questionner sur la signification biologique d'une telle différence, parfois marquée, entre les différents tissus.

Malheureusement, nous n'avons pas pu obtenir les données de protéomique de ces deux 32 tissus, car partagées sous leur forme brute au sein de l'analyse de *Wang et al.* [289]. À l'avenir, il serait intéressant de rajouter les données de protéomique au sein de cette étude pour pouvoir faire le lien entre CUB, ARNm et protéines produites, mais aussi d'approfondir notre catégorisation

Figure 6.1 – **ACP sur le CUB des différents groupes de gènes retrouvés au sein de l'analyse des 32 tissus humains.** Chaque point représente une catégorie de gène pour un des 32 tissus analysés. Les différents tissus de l'analyse sont indiqués par le jeu de couleurs décrit sur la droite de la figure. Ici, seuls les deux premiers axes de l'ACP sont présentés.

	Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Phe	Pro	Ser	Thr	Tyr	Val
adipose tissue	GCU	CGU	GAU	AAC	UGU	GAG	CAG	GGU	CAC	AUC	UUG	AAG	UUC	CCU	UCU	ACC	UAC	GUU
adrenal gland	GCU	CGU	GAU	AAC	UGU	GAA	CAG	GGU	CAC	AUC	UUG	AAG	UUC	CCU	UCU	ACC	UAU	GUU
bone marrow	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAU	AUC	UUG	AAG	UUC	CCC	UCU	ACC	UAC	GUC
cerebral cortex	GCU	CGU	GAU	AAC	UGU	GAA	CAG	GGU	CAU	AUC	UUG	AAG	UUU	CCU	UCU	ACC	UAU	GUU
colon	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCU	UCU	ACC	UAC	GUU
duodenum	GCU	CGU	GAU	AAC	UGC	GAA	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCU	UCU	ACC	UAC	GUU
endometrium	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCA	UCU	ACC	UAU	GUU
esophagus	GCU	CGU	GAU	AAC	UGU	GAA	CAG	GGU	CAU	AUC	UUG	AAG	UUC	CCC	UCU	ACU	UAU	GUU
fallopian tube	GCU	CGU	GAU	AAC	UGC	GAA	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCU	UCU	ACC	UAU	GUU
gall bladder	GCU	CGU	GAC	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCA	UCU	ACC	UAC	GUC
heart	GCU	CGU	GAU	AAC	UGC	GAA	CAG	GGU	CAU	AUC	UUG	AAG	UUU	CCC	UCU	ACC	UAU	GUU
kidney	GCU	CGU	GAU	AAC	UGU	GAA	CAG	GGU	CAC	AUC	UUG	AAG	UUC	CCA	UCU	ACC	UAU	GUU
liver	GCC	CGU	GAC	AAC	UGU	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUC
lung	GCU	CGU	GAC	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUG
lymph node	GCU	CGU	GAC	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUC
ovary	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCU	UCU	ACC	UAC	GUU
pancreas	GCU	CGA	GAU	AAC	UGU	GAA	CAG	GGU	CAU	AUU	CUG	AAG	UUC	CCG	UCU	ACG	UAC	GUA
placenta	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUU	AAG	UUC	CCC	UCU	ACC	UAU	GUU
prostate gland	GCU	CGU	GAU	AAC	UGU	GAA	CAG	GGU	CAC	AUC	UUG	AAG	UUC	CCU	UCU	ACC	UAU	GUU
rectum	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUC
saliva secreting gland	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUC
skeletal muscle tissue	GCU	CGU	GAU	AAC	UGC	GAA	CAA	GGU	CAU	AUU	CUC	AAA	UUU	CCC	UCU	ACU	UAU	GUC
small intestine	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCU	UCU	ACC	UAC	GUU
smooth muscle tissue	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAU	GUU
spleen	GCU	CGU	GAC	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUG
stomach	GCU	CGU	GAU	AAC	UGU	GAA	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCA	UCU	ACU	UAU	GUU
testis	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCU	UCU	ACC	UAC	GUU
thyroid gland	GCU	CGU	GAU	AAC	UGC	GAA	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUU
tonsil	GCU	CGU		AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUC
urinary bladder	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUC
vermiform appendix	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	CUG	AAG	UUC	CCC	UCU	ACC	UAC	GUC
zone of skin	GCU	CGU	GAU	AAC	UGC	GAG	CAG	GGU	CAC	AUC	UUG	AAG	UUC	CCA	UCU	ACC	UAC	GUU

Figure 6.2 – **Codons « optimisés » pour chaque acide aminé parmi les 32 tissus analysés.** La mesure de l'optimisation s'inspire des travaux de *Whittle et Extavour, Whittle et al.*, où le codon synonyme « optimisé » est celui possédant le plus grand écart positif de fréquence entre les gènes fortement exprimés et les autres gènes [39, 58]. Les couleurs jaunes et violettes indiquent les codons synonymes se terminant par AT ou par GC.

des gènes en ajoutant, par exemple, des classes contenant les gènes uniques à certains tissus. En effet, une analyse de *Sémon et al.* démontre qu'il existe une faible variation du CUB au sein des gènes spécifiques à 18 tissus d'*H. sapiens*. Pour *Sémon et al.*, cette variation du CUB s'explique avant tout par la position des gènes étudiés dans le génome, où le contenu en GC des régions codantes et non-codantes sont tous deux influencés par des biais mutationnels, ce qui réfuterait l'existence d'une sélection traductionnelle sur les gènes tissu-spécifiques [287]. Ils indiquent par ailleurs que le CUB partagé entre les orthologues tissu-spécifiques de *H. sapiens* et *M. musculus* ne s'explique pas par une sélection, mais plutôt par une conservation de la position des gènes et de leur environnement génomique. Mais la conservation de la position de ces orthologues pourrait-elle être due à l'environnement nucléotidique de la région concernée ? Pourrait on supposer que ces gènes se trouvent à des endroits spécifiques du génome pour conserver un certain CUB ? Pour répondre à ces questions, il serait nécessaire de faire une analyse extensive des gènes orthologues, de leur CUB et de leur expression tissu-spécifique au sein d'un grand nombre de Vertébrés, et d'observer si la covariation entre le contenu en GC des régions codantes et non-codantes est toujours respectée.

Nous avons notamment produit les résultats précédents pour effectuer une future analyse du CUB des virus d'*H. sapiens* face aux tissus qu'ils infectent. En effet, les virus pourraient représenter un terrain d'expérimentation exceptionnel pour mesurer le CUB chez les Vertébrés. Généralement égoïste à l'échelle cellulaire, certains virus auraient tendance à exploiter la machinerie ribosomale de leur hôte pour augmenter la cadence et la fidélité de production des virions, quitte à écraser dans certains cas celle des protéines de l'hôte. Le CUB des gènes viraux ne fait pas obstacle à de telles observations, où celui-ci, selon l'organisme étudié, peut mimer celui des gènes fortement exprimés au sein de son hôte pour améliorer le processus de traduction de ces propres gènes, ou encore adopter un CUB contraire à son hôte pour baisser sa virulence ou pour éviter la concurrence sur les populations d'ARNt avec les gènes cellulaires fortement exprimés [112, 126, 130, 131]. Mais de telles observations restent souvent limitées aux relation hôte-parasite à certains niveaux taxonomiques, et bien que certaines de ces hypothèses ont pu être explorées à l'échelle des virus humains, celles-ci demeurent encore difficiles à démontrer [130, 290]. Pourtant, l'analyse du CUB des virus pourrait nous aider à déterminer l'existence (ou non) d'une sélection traductionnelle chez les mammifères : si les virus possèdent des signatures claires d'une sélection pour améliorer leur traduction, il serait naturel de penser que la machinerie cellulaire de l'hôte est soumise aux mêmes règles. La vaste diversité des virus, aussi bien sur leur tropisme, sur leur pathogénicité que sur leur style de vie pourrait nous permettre d'explorer cet aspect de la relation hôte-parasite. Les polyomavirus, ou encore les papillomavirus, pourraient parfaitement convenir pour ce genre d'analyse. Ceux-ci semblent posséder une histoire phylogénétique monophylétique, mais possèdent une forte diversité dans leur contenu génomique, dans leur tropisme cellulaire et dans la présentation clinique de l'infection, pour peu qu'il y en ait une (voir chapitre 4) [129]. Comme nous avons pu le voir au travers d'une analyse COUSIN sur les gènes de polyomavirus, ceux-ci possèdent un CUB généralement contraire à celui de leur hôte, et ce quel que soit le gène considéré. En suivant les hypothèses de *Chen et al.*, nous pouvons tout à fait supposer que le CUB des polyomavirus humains est « sélectionné » de manière à ce que la grande majorité des infections soient asymptomatiques [131]. Mais pour aller plus loin, il

serait intéressant d'étudier le possible rôle du CUB dans le tropisme cellulaire des polyomavirus. Plusieurs hypothèses ont été soumises sur le point d'entrée des polyomavirus BKPyV, JCPyV et MCPyV, mais leur affinité avec les différents tissus de l'hôte sont parfois difficiles à cerner [190, 203, 203, 204]. À l'avenir, les données du CUB des différents polyomavirus humains seront comparées aux résultats des analyses que nous avons effectués sur les 32 tissus humains, dans l'optique de déceler une relation entre tropisme cellulaire et CUB de la cellule.

Il reste encore un long chemin à parcourir pour confirmer (ou infirmer) l'existence d'une sélection traductionnelle au sein des Vertébrés. Pour être plus précis, et comme le soulignent *Sémon et al.* au sein de leur article, il est fort probable qu'une sélection traductionnelle soit présente chez ces organismes, mais son impact sur la machinerie cellulaire pourrait être négligeable [287]. Au sein de ce doctorat, nous proposons de nouveaux points de vue et pistes pour explorer le CUB des Vertébrés, et ce dans l'optique de participer à la grande question des mécanismes du CUB chez les organismes.

III

Partie Trois

ANNEXES

A

Annexe Un

ARTICLES PUBLIÉS OU EN COURS DE SOUMISSION

COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences

Jérôme Bourret*, Samuel Alizon, and Ignacio G. Bravo

Centre National de la Recherche Scientifique, Laboratory MIVEGEC (CNRS, IRD, Uni Montpellier), Montpellier, France

*Corresponding author: E-mail: jerome.bourret@ird.fr.

Accepted: November 25, 2019

Abstract

Codon Usage Preferences (CUPrefs) describe the unequal usage of synonymous codons at the gene, chromosome, or genome levels. Numerous indices have been developed to evaluate CUPrefs, either in absolute terms or with respect to a reference. We introduce the normalized index COUSIN (for COdon Usage Similarity INdex), that compares the CUPrefs of a query against those of a reference and normalizes the output over a Null Hypothesis of random codon usage. The added value of COUSIN is to be easily interpreted, both quantitatively and qualitatively. An eponymous software written in Python3 is available for local or online use (http://cousin.ird.fr). This software allows for an easy and complete analysis of CUPrefs via COUSIN, includes seven other indices, and provides additional features such as statistical analyses, clustering, and CUPrefs optimization for gene expression. We illustrate the flexibility of COUSIN and highlight its advantages by analyzing the complete coding sequences of eight divergent genomes. Strikingly, COUSIN captures a bimodal distribution in the CUPrefs of human and chicken genes hitherto unreported with such precision. COUSIN opens new perspectives to uncover CUPrefs specificities in genomes in a practical, informative, and user-friendly way.

Key words: codon usage bias, mutational bias, translational selection, nucleotide composition, amino acid composition, codon adaptation index, bioinformatics, mutation–selection.

Introduction

Translation of messenger RNAs (mRNA) into proteins is a central molecular biology process common to all forms of life. During translation, ribosomes proceed along the mRNA in steps of three nucleotides, called codons. The ribosome allows pairing of a mRNA codon against the complementary anticodon on a transfer RNA (tRNA), catalyzing the polymerization of amino acids to yield peptides and proteins (Quax et al. 2015). Sixty four nucleotide triplets are available and, in the standard genetic code, 61 codons encode for the 20 standard amino acids (Belalov and Lukashev 2013). Because of this asymmetry, certain groups of codons, known as "synonymous codons," encode for the same amino acid (Nirenberg and Matthaei 1961; Khorana et al. 1966). Synonymous codons are not used with similar frequencies, resulting in so-called Codon Usage Preferences (CUPrefs) or Codon Usage bias. Different CUPrefs can be identified in regions within a gene, between genes within a genome and between genomes in different organisms (Grantham et al. 1980; Carbone et al. 2003).

A variety of indices have been developed since the 1980s to describe CUPrefs (Ikemura 1981; Freire-Picos et al. 1994; Urrutia and Hurst 2001; Zhang et al. 2012). Most of them compare the CUPrefs of a query either against a reference set or against a Null Hypothesis (Shields et al. 1988; Lee et al. 2010). The "Codon Adaptation Index" (CAI; Sharp and Li 1987) and the "Effective Number of Codons" (ENC; Wright 1990) are respectively the most popular indices for each category. Numerous software packages to evaluate CUPrefs have been implemented, such as INCA (Supek and Vlahovicek 2004), JCAT (Grote et al. 2005) and CodonW (Peden and Sharp 2005). Most of them compute the CAI, sometimes the ENC, and occasionally other indices (Wan et al. 2004; Angellotti et al. 2007). Still, an important number of indices, such as the scaled χ^2 (Shields et al. 1988) or the "Maximum-likelihood Codon Bias" (MCB; Urrutia and Hurst 2001) cannot be calculated via any dedicated software.

Despite this profusion of alternatives, none of the available indices evaluates CUPrefs simultaneously against a reference and against a Null Hypothesis, thus hindering direct interpretation of the results. We conceived COUSIN (for COdon Usage Similarity INdex) as a score to estimate CUPrefs of a sequence compared with those of a reference, normalized over a Null Hypothesis of equal usage of synonymous codons.

The output is normalized and allows for a straightforward biological interpretation. We have implemented COUSIN together with seven other existing indices in an eponymous Python3 software available for local or online use (http://cousin.ird.fr). To illustrate the power of COUSIN, we compare it to the well-known CAI by analyzing eight complete Coding DNA Sequence (CDSs) data sets from a range of organisms with large differences in nucleotide composition and genome organization.

Measuring CUPrefs with COUSIN

In COUSIN, the CUPrefs of a query are compared with those of a reference data set, and the results of this comparison are normalized over a Null Hypothesis of equal usage of synonymous codons. The notations used as well as the detailed calculation steps are given in table 1.

The amino acid composition of a sequence may affect its CUPrefs (Roth et al. 2012). We therefore conceived two variants of our index: In COUSIN_{18} each of the 18 families of synonymous codons contributes equally to the global index, whereas in COUSIN_{59} each family contributes proportionally to the frequency of the corresponding amino acid in the query. The classical CAI score would thus correspond thus to CAI_{59}. For the sake of comparison we have defined the equivalent CAI_{18} as described in supplementary data 1, Supplementary Material online. By comparing the "18" and "59" scores of an index, we can estimate the impact of amino acid composition on the CUPrefs of a sequence. The COUSIN score calculation involves five steps:

1. Calculate deviation scores ($\text{dev}_{c,a}$) for each codon (c) of each amino acid (a) in the reference data set, compared with the Null Hypothesis:

$$\text{dev}_{c,a} = f_{c,a}^{\text{ref}} - f_{c,a}^{H_0} \quad (1)$$

where $f_{c,a}^{\text{ref}}$ is the frequency of the codon c among its synonymous in the reference data set and $f_{c,a}^{H_0}$ the corresponding frequency under the Null Hypothesis.

2. Define a weight for each codon ($W_{c,a}$), by multiplying the codon frequency in the reference by its deviation score:

$$W_{c,a}^{\text{ref}} = f_{c,a}^{\text{ref}} \times \text{dev}_{c,a} \quad (2)$$

3. Repeat step 2 for the codon frequencies in the query:

$$W_{c,a}^{\text{que}} = f_{c,a}^{\text{que}} \times \text{dev}_{c,a} \quad (3)$$

Using the same deviation score to calculate the weights allows us to compare the scores of the query and of the reference.

Table 1
Notations Used to Define COUSIN and CAI Indices

Symbol	Description
c	Codon
a	Amino acid
f	Frequency
ref	Reference
que	Query
H_0	Null hypothesis
L	Query length
k_a	Set of synonymous codons coding for amino acid a
$\mathcal{A}$	Amino acids present in both query and reference
$\mathcal{N}$	Number of amino acids present in both query and reference

4. The COUSIN_{18}^{a} score of each amino acid is the ratio of the sum of the weights of all synonymous codons for this amino acid in the query data set over the corresponding sum of the weights in the reference data set:

$$\text{COUSIN}_{18}^{a} = \frac{1}{\mathcal{N}} \times \frac{\sum_{c \in k_a} W_{c,a}^{\text{que}}}{\sum_{c \in k_a} W_{c,a}^{\text{ref}}} \quad (4)$$

where $\mathcal{N}$ is the number of amino acids present in both the query and the reference and k_a is the set of synonymous codons coding amino acid a.
For COUSIN_{59}:

$$\text{COUSIN}_{59}^{a} = f_a^{\text{que}} \times \frac{\sum_{c \in k_a} W_{c,a}^{\text{que}}}{\sum_{c \in k_a} W_{c,a}^{\text{ref}}} \quad (5)$$

where f_a^{que} is the frequency of the amino acid a in the query.

5. The final COUSIN score is obtained by adding the individual COUSIN scores of all amino acids:

$$\text{COUSIN}_{18} = \sum_{a \in \mathcal{A}} \text{COUSIN}_{18}^{a} \quad (6)$$

$$\text{COUSIN}_{59} = \sum_{a \in \mathcal{A}} \text{COUSIN}_{59}^{a} \quad (7)$$

By design, the results of COUSIN have an immediate biological interpretation and are directly suitable for hypothesis testing (fig. 1):

- a COUSIN score of 1 indicates that the CUPrefs in the query are similar to those in reference data set;
- a COUSIN score of 0 indicates that the CUPrefs in the query are similar to those in the Null Hypothesis (i.e., equal usage of synonymous codons);
- above 1, CUPrefs in the query are similar to those in the reference but of larger magnitude;

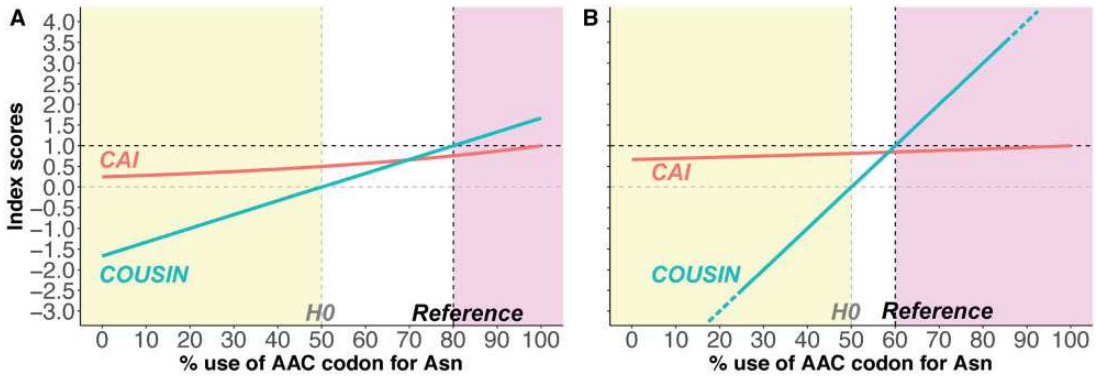

FIG. 1.—COUSIN (blue curve) and CAI (red curve) scores (*y*-axis) for a set of hypothetical queries with different frequency for the AAC and AAU codons encoding the asparagine amino acid (*x*-axis). Values are calculated for a reference set using (*A*) a strong usage bias of AAC:AAU 80:20 and (*B*) a slight usage bias of AAC:AAU 60:40. Vertical dashed lines indicate the composition for the Null Hypothesis of equal usage of both codons (gray line) and for the corresponding reference (black line). Horizontal dashed lines show the COUSIN key values that correspond to the Null Hypothesis (gray line) and to the reference (black line). The yellow area indicates queries with CUPrefs opposite to those in the reference, white one queries with similar but weaker CUPrefs than the reference and pink one queries with similar and stronger CUPrefs than the reference. Notice that, by design, the COUSIN values are always 0 and 1 respectively for the H0 and for the reference, independently of the CUPrefs in the reference. By definition, CAI is bounded by 0 and 1. In this example, COUSIN scores below –3 and above 4 are omitted to facilitate results visualization and reading.

- between 0 and 1, CUPrefs in the query are similar to those in the reference but of smaller magnitude;
- below 0, CUPrefs in the query are opposite to those in the reference;

Upper and lower boundaries to COUSIN values depend on the CUPrefs of the reference: The closer the CUPrefs of the reference are to the null hypothesis, the largest the range of the possible COUSIN scores. As an example, in the case of *Homo sapiens*, with a light global bias in CUPrefs, the range of possible COUSIN values is [–4.48; 6.13]. On the other hand, for *Plasmodium falciparum*, with a strong global bias in CUPrefs, the boundaries for COUSIN values are [0.15; 1.35]. To facilitate interpretation of CUPrefs, artificial boundaries can be given when calculating a COUSIN score. The COUSIN software, described below, proposes such solution.

COUSIN Software

We designed a Python3 software package to implement COUSIN along with other seven existing indices to facilitate CUPrefs analysis and comparisons between methods. The COUSIN software and its documentation are accessible online at http://cousin.ird.fr. A local version can be downloaded from the same website to be used on a UNIX-like Operating System via command lines. For most tasks, the COUSIN software requires query sequences in a FASTA format and a reference data set in a kazusa-like format (Nakamura et al. 2000). The global architecture of the COUSIN software is described in supplementary data 2, Supplementary Material online.

For any entry, the COUSIN software initially calculates basic nucleotide and amino acid composition statistics and estimates CUPrefs. The COUSIN software currently features eight indices that evaluate CUPrefs: COUSIN, CAI (Sharp and Li 1987), ENC (Wright 1990), FOP (Ikemura 1981), SCUO (Angellotti et al. 2007), ICDI (Freire-Picos et al. 1994), CBI (Bennetzen and Hall 1982), and scaled χ^2 (Shields et al. 1988).

If instructed by the user, the COUSIN software performs simulations to assess whether the score of a query is statistically close to that of a standard CDS encoded by the reference (supplementary data 2, Supplementary Material online). The COUSIN software offers additional features to further analyze CUPrefs, such as a clustering analysis to group sequences according to their CUPrefs, or an optimization step to modify the CUPrefs of a sequence to adhere to those in the reference. The COUSIN software can also create a Codon Usage Table in a kazusa-like style from a set of sequences.

COUSIN Analysis

We illustrate the potential of the COUSIN and compare it to the widely used CAI by performing an analysis on the complete CDSs of eight unrelated organisms with contrasted GC content: Two prokaryotes (*Escherichia coli*, *Streptomyces coelicolor*), a plant (*Arabidopsis thaliana*), a yeast (*Saccharomyces cerevisiae*), a protist (*P. falciparum*), a bird (*Gallus gallus*), and two mammals (*H. sapiens*, *Mus musculus*). We extracted the complete nuclear CDSs from these genomes using the Emboss extractfeat function (Rice et al. 2000). To avoid redundancy, when there were alternative spliced forms of a gene, only the first isoform was kept. Only CDSs >300 nucleotides were kept for the analyses. Indeed, most CUPrefs methods show strong biases when analyzing sequences <100 amino acids (Comeron and Aguadé 1998;

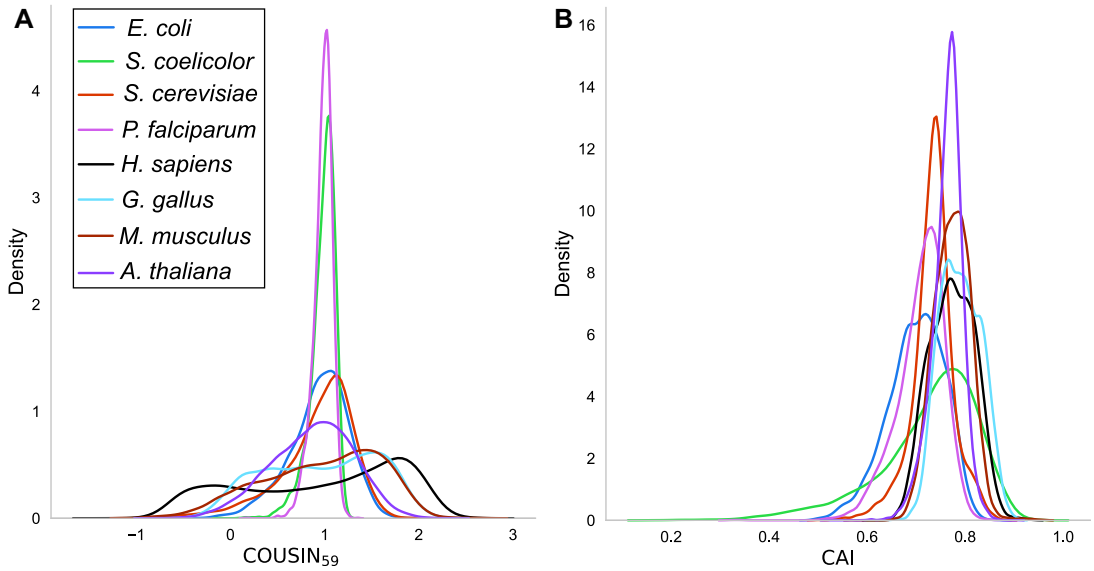

Fig. 2.—Density curves for $COUSIN_{59}$ (*A*) and CAI (*B*) indices for the complete CDSs of the eight organisms studied (see color legend). For each CDS the values for $COUSIN_{59}$ and CAI were calculated against the average codon usage reference table of the corresponding genome. The $COUSIN_{59}$ normalization renders curves centered around 1 allowing for rapid identification of differential dispersion in the leptokurtic curves for organisms with strong nucleotide compositional biases (e.g., *Streptomyces coelicolor*, in green) compared with those more platykurtic for organisms with weaker compositional biases (e.g., *Escherichia coli* in blue). Notice the bimodal distributions for *Homo sapiens* (black) and *Gallus gallus* (light blue) in panel *A*.

Roth et al. 2012). Further details about the selected sequences are in supplementary data 3, Supplementary Material online.

For each genome data set, we constructed a codon usage reference table using the corresponding the COUSIN software utility, and calculated the CUPrefs scores of each CDS against this reference. The resulting density curves for $COUSIN_{59}$ and CAI are presented in figure 2. Details about this analysis are given in supplementary data 4–6, Supplementary Material online.

Analysing CDSs with COUSIN highlights shared patterns as well as differences between organisms. All COUSIN frequency curves (fig. 2, panel A) are centered around 1 (i.e., with similar CUPrefs to those of the reference), but differ strongly in terms of dispersion and of the global shape of data distribution (unimodal, bimodal, or nearly flat). For *S. coelicolor* and *P. falciparum* COUSIN distributions are unimodal and display little variance, consistent with the strong nucleotide composition bias in these genomes (respectively 92.4% GC3 and 17.8% GC3). For other organisms with unimodal distribution but less biased nucleotide composition (e.g., *E. coli*, with 54.9% GC3), the distributions display a larger variance. For larger genomes with strong local differences in nucleotide composition (e.g., chromosome isochores in *H. sapiens*, and microchromosomes in *G. gallus*), the COUSIN frequency curves capture a hitherto not described bimodal shape of CUPrefs (supplementary data 4, Supplementary Material online). For the CAI results obtained with the same data sets (fig. 2, panel B), all frequency curves display unimodal shapes while exhibiting differences in their central value and dispersion, preventing direct contrast with one another.

The key difference between COUSIN and CAI resides in the direct interpretation of the COUSIN results. Indeed, the correlation between CAI and COUSIN scores for each CDS is strong and positive, ranging from 0.661 in *A. thaliana* to 0.978 in *S. coelicolor* (see complete comparisons in supplementary data 6, Supplementary Material online). However, for COUSIN we compare here the CUPrefs of individuals CDSs to a reference representing the average CUPrefs of the organism, therefore expecting—and obtaining—an average score close to 1. For CAI, the central value of the obtained distribution depends on the precise CUPrefs of the reference, and are therefore not comparable between organisms. This lack of normalization hampers any direct comparison of CAI values for genes against different reference sets. Furthermore, the COUSIN score seems to better capture the impact of the query's GC3 content on CUPrefs with, for instance, a Pearson correlation score of 0.91 between GC3 and $COUSIN_{59}$ and of 0.86 between GC3 and CAI for *H. sapiens* CDSs (supplementary data 4, Supplementary Material online).

Discussion

A large number of indices have been conceived to evaluate CUPrefs. Nevertheless, in most cases they do not allow for a straightforward interpretation. As an example, the CAI is often considered as a direct measure of CUPrefs against a reference. However, the CAI value of the reference against itself

is different for each reference, preventing comparisons between genomes. Further, the CAI value of 1 is virtually never reached by any CDS in a given genome. Similarly, the different flavors of ENC (Wright 1990; Novembre 2002; Satapathy et al. 2017) allow to evaluate the presence and extent of CUPrefs against a Null Hypothesis of equal codon usage, but cannot inform on the precise trends of the detected CUPrefs.

We introduced here COUSIN, a new index to measure the CUPrefs of a sequence with respect to both a reference and a Null Hypothesis of equal usage of synonymous codons. The COUSIN value has a straightforward quantitative and qualitative meaning: It allows for an easy comparison 1) between the CUPrefs of the query CDS and those of both the reference and random CUPrefs, and 2) between queries and/or between data sets. We implemented the calculation of the COUSIN index, as well as of a number of additional features and existing indices to evaluate CUPrefs, into an eponymous bioinformatic software, available in a stand-alone as well as in an online version (COUSIN, at http://cousin.ird.fr).

We briefly illustrated the novelty and potential of the COUSIN by analyzing all CDSs in the genomes of eight divergent organisms. Taking the average genomic CUPrefs as a reference, we showed that COUSIN brings to light strong differences between CDSs within organisms, as well as between organisms. Such differences are far less obvious when using the CAI. Importantly, using the average genomic CUPrefs as a reference may or not be relevant when analyzing tissue or condition-dependent CUPrefs based on gene expression data. It remains the responsibility of the user to choose the appropriate reference and to interpret the results accordingly. Our results on differences in COUSIN values distribution and variance (exemplified by the bimodality in *H. sapiens* and *G. gallus*) demonstrate the power and utility of this novel index to identify differential heterogeneity between and within genomic data sets.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

J.B. is funded by a PhD fellowship from the French Ministry of Education and Research. This study was supported by the European Union's Horizon 2020 research and innovation program under the grant agreement CODOVIREVOL (ERC-2014-CoG-647916) to I.G.B. The authors acknowledge the CNRS and the IRD for additional, intramural support. The computational results presented have been achieved in part using the IRD Bioinformatic Cluster *itrop*, which also hosts the COUSIN online server (http://cousin.ird.fr). We thank Frédéric Delsuc for driving our attention onto the composition particularities of the *G. gallus* genome.

Literature Cited

Angellotti MC, Bhuiyan SB, Chen G, Wan XF. 2007. CodonO: codon usage bias analysis within and across genomes. Nucleic Acids Res. 35(Web Server):W132–W136.
Belalov IS, Lukashev AN. 2013. Causes and implications of codon usage bias in RNA viruses. PLoS One 8(2):e56642.
Bennetzen JL, Hall BD. 1982. Codon selection in yeast. J Biol Chem. 257(6):3026–3031.
Carbone A, Zinovyev A, Képès F. 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19(16):2005–2015.
Comeron JM, Aguadé M. 1998. An evaluation of measures of synonymous codon usage bias. J Mol Evol. 47(3):268–274.
Freire-Picos MA, et al. 1994. Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes. Gene 139(1):43–49.
Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8(1):r49–r62.
Grote A, et al. 2005. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res. 33(Web Server):W526–W531.
Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol. 151(3):389–409.
Khorana HG, et al. 1966. Polynucleotide synthesis and the genetic code. Cold Spring Harb Symp Quant Biol. 31:39–49.
Lee S, Weon S, Lee S, Kang C. 2010. Relative codon adaptation index, a sensitive measure of codon usage bias. Evol Bioinform Online. 6:47–55.
Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 28(1):292.
Nirenberg MW, Matthaei JH. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A. 47(10):1588–1602.
Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol. 19(8):1390–1394.
Peden J, Sharp P. 2005. CodonW: Correspondence analysis of Codon Usage. http://codonw.sourceforge.net/; last accessed November 2019.
Quax TEF, Claassens NJ, Söll D, vanderOost J. 2015. Codon bias as a means to fine-tune gene expression. Mol Cell. 59(2):149–161.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16(6):276–277.
Roth A, Anisimova M, Cannarozzi G. 2012. Measuring codon usage bias. In Codon Evolution Mechanisms and Models, Chapter: 13. Oxford University Press.
Satapathy SS, Sahoo AK, Ray SK, Ghosh TC. 2017. Codon degeneracy and amino acid abundance influence the measures of codon usage bias: improved Nc (Nc) and ENCprime (N'c) measures. Genes Cells 22(3):277–283.
Sharp PM, Li WH. 1987. The codon Adaptation Index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15(3):1281–1295.
Shields DC, Sharp PM, Higgins DG, Wright F. 1988. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol. 5(6):704–716.
Supek F, Vlahovicek K. 2004. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics 20(14):2329–2330.

Urrutia AO, Hurst LD. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159(3):1191–1199.

Wan XF, Xu D, Kleinhofs A, Zhou J. 2004. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evol Biol. 4(1):19.

Wright F. 1990. The 'effective number of codons' used in a gene. Gene 87(1):23–29.

Zhang Z, et al. 2012. Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. BMC Bioinformatics 13(1):43.

Associate editor: Gwenael Piganeau

EVOLUTION OF DIFFERENTIAL CODON USAGE PREFERENCES AND SUBFUNCTIONALISATION IN PARALOGOUS GENES: THE SHOWCASE OF POLYPYRIMIDINE TRACT BINDING PROTEINS

Jérôme Bourret[1, †, *], Fanni Borvető[1, †], and Ignacio G. Bravo[1]

[1]Centre National de la Recherche Scientifique (CNRS), Laboratoire MIVEGEC (CRNS IRD UM), Montpellier, France

[†]These authors contributed equally to this work

ABSTRACT

Gene paralogs are copies of a same gene that appear after gene or full genome duplication. Redundancy generated by gene duplication may release certain evolutionary pressures, allowing one of the copies to access novel gene functions. Here we focused on role of codon usage preferences (CUPrefs) during the evolution of the polypyrimidine tract binding protein (*PTBP*) splicing regulator paralogs.

PTBP1-3 show high identity at the amino acid level (up to 80%), but display different nucleotide composition, divergent CUPrefs and distinct tissue-specific expression levels. Phylogenetic inference differentiates the three orthologs and suggests that the three *PTBP1-3* lineages predate the basal diversification within vertebrates. We identify a distinct substitution pattern towards GC3-enriching mutations in *PTBP1*, with a trend for the use of common codons and for a tissue-wide expression. Genomic context analysis shows that GC3-rich nucleotide composition for *PTBP1*s is driven by local mutational processes. In contrast, *PTBP2*s are enriched in AT-ending, rare codons, and display tissue-restricted expression. Nucleotide composition and CUPrefs of *PTBP2* are only partly driven by local mutational forces, and could have been shaped by selective forces. Interestingly, trends for use of UUG-Leu codon match those of AT-ending codons.

Our interpretation is that a combination of mutation and selection has differentially shaped CUPrefs of *PTBPs* in Vertebrates: GC-enrichment of *PTBP1* is linked to the strong and broad tissue-expression, while AT-enrichment of *PTBP2* and *PTBP3* is linked to rare CUPrefs and specialized spatio-temporal expression. Our model is compatible with a gene subfunctionalisation process by differential expression regulation associated to the evolution of specific CUPrefs.

Keywords Codon usage bias, codon usage preferences, gene duplication, paralog, ortholog, evolution, mutation-selection, nucleotide composition, tissue-specific expression

[*]Corresponding author. email : jerome.bourret@ird.fr

1 Introduction

During translation, ribosomes assemble proteins by specific amino acid linear polymerisation guided by the successive reading of mRNA nucleotide triplets called codons. Each time a codon is read, it is chemically compared to the set of available tRNAs' anticodons. Upon codon-anticodon sequence match the ribosome loads the tRNA and adds the associated amino acid to the nascent protein. The main 20 amino acids are decoded by 61 codon-anticodon combinations, so that multiple codons are associated to the same amino acid and are named synonymous codons (Nirenberg and Matthaei, 1961; Khorana et al., 1966). Codon Usage Preferences (CUPrefs) refer to the differential usage of synonymous codons, between species, or between genes and genomic regions in the same genome (Grantham et al., 1980; Carbone et al., 2003). Mutation and selection are the two main forces shaping CUPrefs (Duret, 2002; Chamary et al., 2006; Plotkin and Kudla, 2011). Mutational biases relate to directional mechanistic biases during genome replication (Reijns et al., 2015; Apostolou-Karampelis et al., 2016), during genome repair (Lujan et al., 2012), or during recombination (Pouyet et al., 2017), preferentially introducing one nucleotide over others or inducing recombination and maintaining genomic regions depending on their composition. Mutational biases are well known in prokaryotes and eukaryotes, ranging from simple molecular preferences towards 3'A-ending in the *Taq* polymerase (Clark, 1988) to the complex GC-biased gene conversion in vertebrates (Pouyet et al., 2017). Selective forces shaping CUPrefs are often described as translational selection. This notion refers to the ensemble of mechanistic steps and interactions during translation that are affected by the particular CUPrefs of the mRNA, so that the choice of certain codons at certain positions may actually enhance the translation process and can be subject to selection (Bulmer, 1991). Translational selection covers thus codon-mediated effects acting on mRNA maturation, secondary structure and overall stability (Presnyak et al., 2015; Novoa and Ribas de Pouplana, 2012), subcellular localisation, programmed frameshifts, translation speed and accuracy, or protein folding (Caliskan et al., 2015; Mordstein et al., 2020; Spencer and Barral, 2012). Translational selection has been demonstrated in unicellular prokaryotes and eukaryotes (Satapathy et al., 2016; Percudani et al., 1997; Duret and Mouchiroud, 1999; Whittle and Extavour, 2016), often in the context of tRNA availability (Ikemura, 1981). However, its very existence in multi-cellular eukaryotes remains highly debated (Pouyet et al., 2017; Galtier et al., 2018).

Homologous genes share a common origin either by speciation (orthology) or by duplication events (paralogy) (Sonnhammer and Koonin, 2002). Upon gene (or full genome) duplication, the new genome will contain two copies of the original gene, referred to as in-paralogs. After speciation, each daughter cell will inherit one couple of paralogs, *i.e.* one copy of each ortholog (Koonin, 2005). The emergence of paralogs by gene duplication releases the evolutionary constraints on the individual genes. Evolution can thus potentially lead to function specialisation, such as evolving a particular substrate preferences, or engaging each paralog on specific enzyme activity preferences in the case of promiscuous enzymes (Copley, 2020). Gene duplication can also allow one paralog to explore broader sequence space and to evolve radically novel functions, while the remaining counterpart can assure the original function.

The starting point for our research are the experimental observations by Robinson and coworkers reporting differential expression of the polypyrimidine tract binding protein (*PTBP*) human paralogs as a function of their nucleotide com-

position (Robinson et al., 2008). Vertebrates genomes encode for three in-paralogous versions of the *PTBP* genes, all of them fulfilling similar functions in the cell: they form a class of hnRNP RNA-Binding Proteins that are involved in the modulation of mRNAs alternative splicing (Pina et al., 2018). Within the same genome the three paralogs display high amino-acid sequence similarity, around 70% in humans and with similar overall values in vertebrates (Pina et al., 2018).

Despite the high resemblance at the protein level, the three *PTBP* paralogs sharply differ in nucleotide composition, CUPrefs and tissue expression pattern. In humans, *PTBP1* is enriched in GC-ending synonymous codons and is widely expressed in all tissues, while *PTBP2* and *PTBP3* are AT3-rich and display an enhanced expression in the brain and in hematopoietic cells respectively (Supplementary Material S1). Robinson and coworkers studied the expression in human cells of all three human *PTBP* paralogous genes placed under the control of the same promoter. They showed that the GC-rich paralog *PTBP1* was more highly expressed than the AT-rich ones, and that the expression of the AT-rich paralog *PTBP2* could be enhanced by synonymous codons recoding towards the use of GC-rich codons (Robinson et al., 2008). Here we have built on the evolutionary foundations of this observation and extended the analyses of CUPrefs to *PTBP* paralogs to vertebrate genomes. Our results suggest that paralog-specific directional changes in CUPrefs in mammalian *PTBP* concurred with a process of subfunctionalisation by differential tissue pattern expression of the three paralogous genes.

2 Material and Methods

Sequence retrieval

We assembled a dataset of DNA sequences from 47 mammalians and 27 non-mammalians Vertebrates and 3 protostomes using the BLAST function on the nucleotide database of NCBI (NCBI Resource Coordinators, 2018) using the human *PTBP* paralogs as references (see supplementary Material S2 for accession numbers). We could identify the corresponding three ortholog genes in all Vertebrates species screened except for the European rabbit *Oryctolagus cuniculus*, lacking *PTBP1* and from the rifleman bird *Acanthisitta chloris*, lacking *PTBP3* (Supplementary Material S2). The final vertebrate dataset contained 75 *PTBP1*, 76 *PTBP2* and 75 *PTBP3* sequences. As outgroups for the analysis, we retrieved the orthologous genes in three protostomes genomes, which contained a single *PTBP* homolog per genome (Supplementary Material S3). From the original dataset, we identified a subset of nine mammalian and six non-mammalian vertebrates species with a good annotation of the *PTBP* chromosome context, and we retrieved compositional information on the flanking regions and on the intron composition (Supplementary Material S3). Because of annotation hazards, intronic and flanking regions information were missing for some *PTBP*s in the African elephant *Loxodonta africana*, Schlegel's Japanese Gecko *Gekko japonicus* and the whale shark *Rhincodon typus* assemblies. For these 15 species the values for codon adaptation index (CAI) (Sharp and Li, 1987) and codon usage similarity index (COUSIN) (Bourret et al., 2019) were calculated using the COUSIN server (available at https://cousin.ird.fr).

Clustering PTBPs by their CUPrefs

For each *PTBP* paralog we calculated codon composition and CUPrefs analyses via the COUSIN tool (Bourret et al., 2019). For each *PTBP* gene we constructed a vector of 59 positions with the relative frequencies of all synonymous codons. As tools for information dimension reduction to analysis CUPrefs we applied on the 229 59-dimension vectors: i) a k-means clustering; ii) a hierarchical clustering; and iii) a principal component analysis (PCA).

Alignment and phylogenetic analyses

To generate robust alignments without introducing artefacts due to large evolutionary distances between in-paralogs we proceeded stepwise, as follows: i) we aligned separately at the amino acid level each set of *PTBP* paralog sequences of mammals and non-mammalians Vertebrates; ii) for each *PTBP* paralog we merged the alignments for mammals and for non mammals, obtaining the three *PTBP1*, *PTBP2* and *PTBP3* alignments for all Vertebrates; iii) we combined the three alignments for each paralog into a single one; iv) we aligned the outgroup sequences to the global Vertebrate *PTBP*s alignment. All alignments steps were performed using MAFFT (Katoh et al., 2002). The final amino acid alignment was back-translated to obtain the codon-based nucleotide alignment. The codon-based alignment was trimmed using Gblocks (Castresana, 2000).

Phylogenetic inference was performed at the amino acid and at the nucleotide level using RAxML v8.2.9 and bootstrapping over 1000 cycles (Stamatakis, 2014). For nucleotides we used codon-based partitions and applied the GTR+G4 model while for amino acids we applied the LG+G4 model. For the 79 species used in the analyses we retrieved a species-tree from the TimeTree tool (Kumar et al., 2017). Distances between phylogenetic trees were computed using the Robinson-Foulds index, which accounts for differences in topology (Robinson and Foulds, 1981), and the K-tree score, which accounts for differences in topology and in branch length (Soria-Carrasco et al., 2007). After phylogenetic inference we computed marginal ancestral states for the respectively most recent common ancestors at the nucleotide level of each paralog using RAxML. Using these ancestral sequences we estimated the number of synonymous and non-synonymous mutations of each extant sequence to the corresponding most recent common ancestor.

Statistical analyses

Correlation between matrices was assessed via the Mantel test. Non-parametric comparisons were performed using the Wilcoxon-Mann-Whitney test for population medians and the Wilcoxon signed rank test for paired comparisons. Statistical analyses were performed using the *ape* and *ade4* R packages and JMP v14.3.0.

3 Results

Vertebrate PTBP paralogs differ in nucleotide composition

In order to understand the evolutionary history of *PTBP* genes we performed first a nucleotide composition and CUPrefs analysis on the three paralogs in 79 species. Overall, *PTBP1* are GC-richer than *PTBP2* and *PTBP3* (respective mean percentages 55.9, 42.3 and 44.9 for GC content and 69.5, 33.4 and 38.3 for GC3 content; Figure 1, Supplementary Material S2). In addition, *PTBP1* show a difference in GC3 between mammalian and non-mammalian gene (respectively 79.8 against 59.9 mean percentages). A linear regression model followed by a Tukey's honest significant differences analysis for GC3 using as explanatory levels paralog (*i.e. PTBP1-3*), taxonomy (*i.e.* mammalian or non-mammalian) and their interaction identifies three mains groups of *PTBP*s (Table 1): a first one corresponding to mammalian *PTBP1*, a second one grouping non-mammalian *PTBP1* and a third one spanning all *PTBP2* and *PTBP3*. The largest explanatory factor for GC3 was the paralog *PTBP1-3*, accounting alone for 65% of the variance, while the interaction between the levels taxonomy and paralog captured around 15% of the remaining variance (Table 1). These trends are confirmed when performing paired comparisons between paralogs present in the same mammalian genome, with significant differences in GC3 content in the following order: *PTBP1* > *PTBP3* > *PTBP2* (Wilcoxon

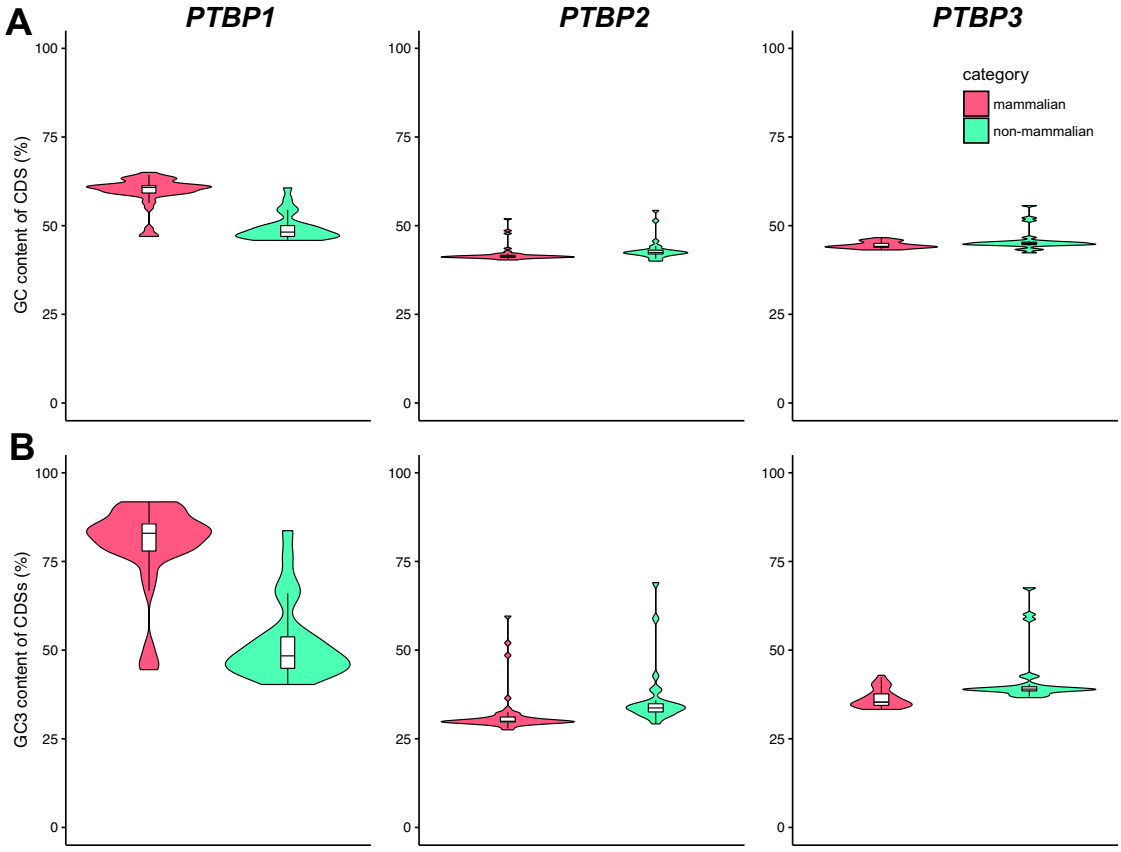

Figure 1: **GC content (A) and GC3 content (B) of Vertebrates *PTBP*s.** Violin plots display the overall distribution while box and whiskers display median, quartiles and 95% of the corresponding values for mammalian (red) and non-mammalian (blue) individual genomes.

signed rank test: *PTBP1* vs *PTBP2*, mean diff=48.0, S=539.50, p-value <0.0001; *PTBP1* vs *PTBP3*, mean diff=43.5, S=517.50, p-value <0.0001; *PTBP3* vs *PTBP2*, mean diff=4.5, S=406.50, p-value <0.0001). Note that even if all of them significantly different, the mean paired differences in GC3 between *PTBP1* and *PTBP2-3* are ten times larger than the corresponding mean paired differences between *PTBP2* and *PTBP3*.

The distribution of the residuals between observed and expected values after our model fit to the data allows to identify a number of outliers species with interesting taxonomical patterns in compositional deviation (Table 2). For non mammals, the three *PTBP* paralogs in the rainbow trout *Oncorhynchus mykiss* genome display high GC3 content (between 67% and 76%), all of them significantly higher than model-predicted values (expected values between 36% and 51%). A similar case occurs for the zebrafish *Danio rerio* genome: the three paralogs display GC3 values around 58%, which for *PTBP2* and *PTBP3* paralogs is significantly higher than predicted by the model (expected values around 38%). Very interestingly, for the monotrema platypus *Ornithorhynchus anatinus* as well as for the three marsupials in the dataset the Tasmanian devil *Sarcophilus harrisii*, the koala *Phascolarctos cinereus* and the grey short-tailed opossum *Monodelphis domestica* their *PTBP1* genes present similar GC3 content around 47%, which is significantly lower than predicted by the model (expected values around 79%).

In many vertebrate species, strong compositional heterogeneities are observed along chromosomes often referred to as "isochores". To explore the influence of the genomic environment on the nucleotide composition of *PTBP*s, for 15 species with well-annotated genomes we analyzed the correlation of paralog GC3 with two local compositional variables of the corresponding gene (GC content of intronic and flanking regions) and with three global compositional variables for the corresponding genomes (global GC3 in the complete genomic ORFome, global GC content in all introns, and global GC content in all flanking regions) (Table 3 and Figure 2). First, for *D. rerio* the GC3 composition of *PTBP2* and *PTBP3* is clearly different from the rest, in line with the outlier results presented in Table 2. We have thus excluded the zebra fish values and performed an individual as well as a stepwise linear fit to explain the variance in GC3 composition by the variance in the local and global compositional variables mentioned above (Table 3). For all three *PTBP*s the local GC content explains best the corresponding GC3 content, but with strong differences between paralogs: while variation in the local composition captures almost perfectly variation in the GC3 content in the case of *PTBP1* (R^2=0.97) and strongly in the case of for *PTBP3* (R^2=0.78), the fraction of variance explained by the local composition significantly drops in the case of *PTBP2* (R^2=0.46).

Vertebrate *PTBP* paralogs differ in CUPrefs

For each *PTBP* coding sequence we extracted the relative frequencies of synonymous codons and performed different approaches to reduce information dimension and visualise CUPrefs trends. The results of a principal component analysis (PCA) are shown in Figure 3. The first PCA axis captured 68.9% of the variance, far before the second and

Table 1: Global linear regression model and post-hoc Tukey's honest significant differences (HSD) test for GC3 composition as explained variable and the explanatory levels paralog (*PTBP1-3*), taxonomy (*i.e.* mammalian or non-mammalian) and their interactions. Overall goodness of the fit: Adj Rsquare=0.83; F ratio=205.7; Prob > F: <0.0001.Individual effects for the levels: i) paralog: F ratio=274.3; Prob > F: <0.0001; ii) taxonomy: F ratio=27.2; Prob > F: <0.0001; iii) interaction paralog*taxonomy: F ratio=87.9; Prob > F: <0.0001.

Level	**Least Sq. Mean (GC3%)**	**Standard error**	**Tukey's HSD group**
Paralog			
PTBP1	65.87	1.00	A
PTBP3	39.00	1.01	B
PTBP2	34.03	1.00	C
Taxonomy			
mammalian	49.32	0.70	A
non-mammalian	43.28	0.92	B
Paralog*Taxonomy			
PTBP1, mammalian	79.81	1.22	A
PTBP1, non-mammalian	51.93	1.59	B
PTBP3, non-mammalian	41.64	1.62	C
PTBP3, mammalian	36.36	1.22	C, D
PTBP2, non-mammalian	36.27	1.59	C, D
PTBP2, mammalian	31.79	1.20	D

the third axes (respectively 6.7% and 3.2%). In codon families with multiplicity two, the two codons are necessarily symmetrically related in the PCA, creating a redundancy. We thus simplified the analysis by performing again a PCA using only the codon families of multiplicity four and six, obtaining similar results (Supplementary Material S5 B). Codons segregate in the first axis by their GC3 composition, the only exception being the UUG-Leu codon, which grouped together with AT-ending codons. This first axis differentiates mammalian *PTBP1*s on the one hand and *PTBP2*s and *PTBP3*s on the other hand. Non-mammalian *PTBP1*s scatter between mammalian *PTBP1*s and *PTBP3*s, along with the protostoma *PTBP*s. In the second PCA axis the only obvious (but nevertheless cryptic) codon-structure trends are: i) the split between C-ending and G-ending codons, but not between A-ending and U-ending codons; and ii) the large contribution in opposite directions to this second axis of the AGA and AGG-Arginine codons. This second PCA axis differentiates *PTBP2*s from *PTBP3*s paralogs, consistent with these composition trends, a paired-comparison confirms that *PTBP3*s are richer in C-ending codons than *PTBP2*s, respectively 21.7% against 15.4% (Wilcoxon signed rank test: mean diff=6.2, S=1184.0, p-value <0.0001).

As an additional way to identify groups of genes with similar CUPrefs we applied a hierarchical clustering and a k-means clustering. Both analyses mainly aggregate *PTBP* genes by their GC3 richness. The *PTBP* dendrogram resulting of the hierarchical clustering (rows in clustering in Figure 3) shows five main clades that cluster the paralogs with a good match to the following groups: mammalian *PTBP1*s, non-mammalian *PTBP1*s, *PTBP2*s, *PTBP3*s and a fifth group containing the protostomata *PTBP*s and a few individuals of all three paralogs (Kappa-Fleiss consistency score = 0.76). Regarding codon clustering, the hierarchical stratification sharply splits GC-ending codons from AT-ending codons, with the only exception again of the UUG-Leu codon, which consistently groups within the AT-ending

Table 2: Individual genes with outlier values with respect to the linear regression expected values for the levels paralog (*PTBP1-3*), taxonomy (mammalian or non-mammalian) and their interactions.

Species	**paralog**	**observed GC3 (%)**	**expected GC3 (%)**	**deviation GC3 (%)**
mammalian				
Desmodus rotundus	*PTBP2*	59.60	31.79	27.81
Miniopterus natalensis	*PTBP2*	48.52	31.79	16.72
Monodelphis domestica	*PTBP1*	44.49	79.81	-35.32
Ornithorhynchus anatinus	*PTBP1*	51.14	79.81	-28.67
Ornithorhynchus anatinus	*PTBP2*	52.00	31.79	20.21
Phascolarctos cinereus	*PTBP1*	47.53	79.81	-32.28
Sarcophilus harrisii	*PTBP1*	45.44	79.81	-34.37
non-mammalian				
Danio rerio	*PTBP2*	58.89	36.27	22.62
Danio rerio	*PTBP3*	60.08	41.64	18.44
Lepisosteus oculatus	*PTBP3*	58.73	41.64	17.10
Oncorhynchus mykiss	*PTBP1*	76.27	51.93	24.34
Oncorhynchus mykiss	*PTBP2*	69.03	36.27	32.76
Oncorhynchus mykiss	*PTBP3*	67.58	41.64	25.95
Pogona vitticeps	*PTBP1*	83.68	51.93	31.75

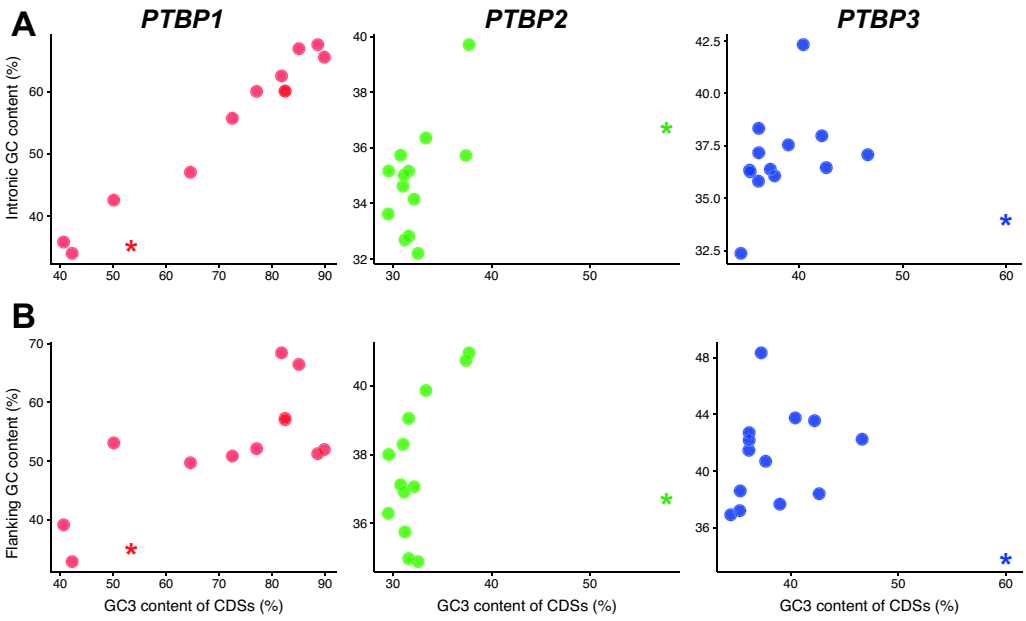

Figure 2: **Variation in GC3 content of *PTBP*s (x-axis) and in the GC content of the corresponding introns (A, y axis) or flanking regions (B, y axis)**. Each dot represents one of the 15 individual used for the genomic context analysis. The asterisk indicates the values for the species *Danio rerio*, which shows peculiar results for *PTBP2* and *PTBP3*, consistent with its outlier behaviour in the global model.

codons. The elbow approach of k-means clustering identifies an optimal number of four clusters and separates the paralog genes with a good match as following: *PTBP1*, *PTBP2*, *PTBP3* and a group containing the protostoma and individuals from all paralogs (Kappa-Fleiss consistency score = 0.75).

Overall, k-means clustering and hierarchical clustering, both based on the 59-dimensions vectors of the CUPrefs, are congruent with one another (Kappa-Fleiss consistency score = 0.83), and largely concordant with the PCA results. CUPrefs define thus groups of *PTBP* genes consistent with their orthology and taxonomy. It is interesting to note that for some species the *PTBP* paralogs display unique distributions of CUPrefs, such as an overall similar CUPrefs in the three *PTBP* genes of the whale shark *Rhincodon typus*, or again some shifts in nucleotide composition between paralogs in the Natal long-fingered bat *Miniopterus natalensis*.

In order to characterise the directional CUPrefs bias of the different paralogs, we have analysed for the 15 species with well-annotated genomes described above, the match between each individual *PTBP* and the average CUPrefs of the corresponding genome (Table 4). Our results highlight strong differences for mammalian paralogs: *PTBP1*s display COUSIN values above 1 while *PTBP2*s display COUSIN values below zero. Given the interpretation of COUSIN values (Bourret et al., 2019) these results mean that in mammals *PTBP1*s are enriched in commonly used codons in a higher proportion that the average in the genome, while *PTBP2*s are enriched in rare codons so that their CUPrefs go in the opposite direction to the average in the genome.

Phylogenetic reconstruction of *PTBP*s

We explored the evolutionary relationships between *PTBP*s by phylogenetic inference at the amino acid and at the nucleotide level (4, Supplementary Material S?). Our final dataset contained 74 *PTBP* sequences from mammals (47 species within 39 families) and non mammal vertebrates (27 species within 24 families). We used the *PTBP* genes from three protostome species as outgroups. Both amino acid and nucleotide phylogenies rendered three main clades grouping the *PTBP*s by orthology. In both topologies, *PTBP1* and *PTBP3* orthologs cluster together, although the protostome outgroups are linked to the tree by very a long branch making it difficult the proper identification of the Vertebrate *PTBP* tree root. Amino acid and nucleotide subtrees are largely congruent (see topology and branch length comparisons in Table5). The apparently large nodal and split distance values between nucleotide and amino acid *PTBP2* trees stem from disagreements in very short branches, as evidenced by the lowest K-tree score for this ortholog

Table 3: Results for an individual or for a sequential least squares regression for explaining variation in GC3 composition of *PTBP*s genes, by variation of different local or of global compositional variables in 14 well-annotated vertebrate genomes. For each gene, individual variables are ordered according to their contribution to the sequentially better model. Variables labelled with N.S. (not significant) do not contribute with significant additional explanatory power when added to the sequential model. BIC, Bayesian information content.

PTBP1				
	Individual contribution		**Sequential contribution**	
Parameter	**R^2**	**BIC**	**R^2**	**BIC**
Local intronic GC	0.96	74.42	0.96	74.42
Global intronic GC	0.03	111.98	0.97	71.23
Global flanking GC	0.05	111.70	0.98 (N.S.)	72.26
Global exomic GC3	0.62	100.71	0.98 (N.S.)	74.27
Local flanking GC	0.55	112.66	0.98 (N.S.)	76.55
PTBP2				
	Individual contribution		**Sequential contribution**	
Parameter	**R^2**	**BIC**	**R^2**	**BIC**
Local flanking GC	0.46	60.12	0.46	60.12
Global flanking GC	0.03	67.66	0.49 (N.S.)	61.86
Local intronic GC	0.37	61.95	0.49 (N.S.)	64.38
Global exomic GC3	0.09	66.75	0.49 (N.S.)	66.89
Global intronic GC	0.05	67.38	0.50 (N.S.)	69.35
PTBP3				
	Individual contribution		**Sequential contribution**	
Parameter	**R^2**	**BIC**	**R^2**	**BIC**
Local intronic GC	0.78	78.11	0.78	78.11
Global intronic GC	0.12	96.38	0.80 (N.S.)	79.56
Global exomic GC3	0.02	97.73	0.82 (N.S.)	80.66
Local flanking GC	0.38	91.77	0.84 (N.S.)	81.70
Global flanking GC	0.02	97.77	0.84 (N.S.)	84.27

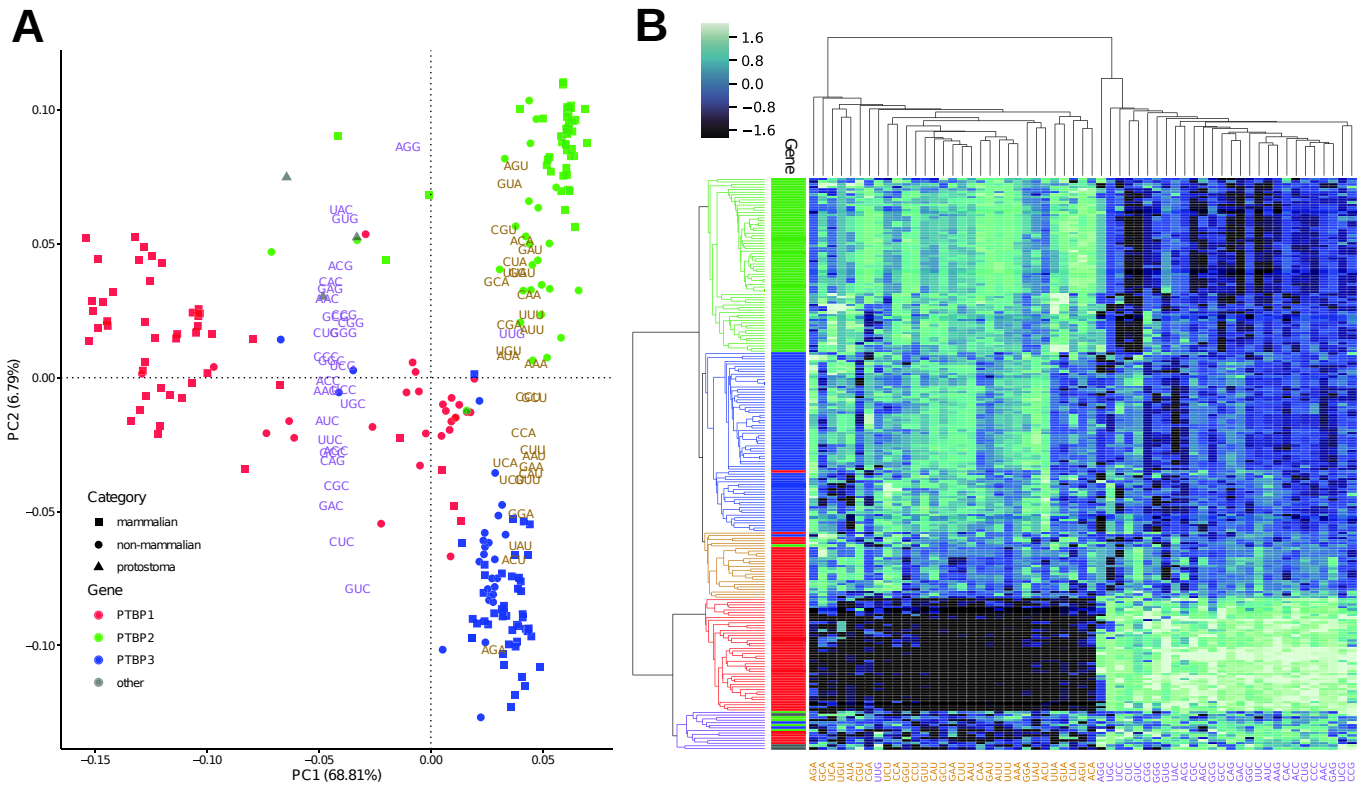

Figure 3: **CUPrefs analysis of *PTBP*s**. A) Plot of the two first dimensions of a PCA analysis based on the codon usage preferences of *PTBP1*s (red), *PTBP2*s (green), *PTBP3*s (blue) and protostoma (grey) individuals. Taxonomic information is included as mammals (squares), non-mammals (circles) and protostomates (triangles). The PCA was created using as variables the vectors of 59 positions (representing the relative frequencies of the 59 synonymous codons) for each individual gene. The eigenvalues of the individual codon variables are given by their position on the graph. Each codon variable is identified by its name and by a colour code, purple for GC-ending codons and orange for AT-ending codons. The percentage of the total variance explained by each axis is shown in parenthesis. B) Heatmap of *PTBP*s individuals (rows) and synonymous codons (columns). Left dendrogram represents the hierarchical clustering of *PTBP*s based on their CUPrefs with colour codes that stand for the clusters created from this analysis. Side bars give information on heatmap individuals regarding i) their origin : *PTBP1* (red), *PTBP2* (green), *PTBP3* (blue) or protostoma (grey). Note the position of the UUG-Leu codon, in both the PCA and the codon dendrogram, as the sole GC-ending codon clustering with all other AT-ending codons)

(as a reminder, the Robinson-Foulds index exclusively regards topology while the K-tree score combines topological and branch-length dependent distance between trees, see Material and Methods). In all three cases, internal structure of the ortholog trees essentially recapitulates species taxonomy at the higher levels (Table5). Some of the species identified by the mathematical model as displaying a largely divergent nucleotide composition present accordingly long branches in the phylogenetic reconstruction, such as *PTBP3* for *O. mykiss*.

We have then analysed the correspondence between nucleotide-based and amino acid-based pairwise distances. We observe a good correlation between both reconstructions for all paralogs, except for mammalian *PTBP2*s, which display

217 extremely low divergence at the amino acid level (Figure 5 B, Supplementary Material S8 B). For mammalian *PTBP1*s,
218 the plot allows to clearly differentiate a cloud with the values corresponding to the monotremes+marsupial mammals,

Table 4: Global linear regression model and post-hoc Tukey's honest significant differences (HSD) test, the explained variable being the COUSIN value of the each *PTBP* gene against the average of the corresponding genome and the explanatory levels paralog (*PTBP1-3*), taxonomy (*i.e.* mammalian or non-mammalian) and their interactions. Overall goodness of the fit: Adj Rsquare=0.82; F ratio=36.84; Prob > F: <0.0001.Individual effects for the levels: i) paralog: F ratio=40.72; Prob > F: <0.0001; ii) taxonomy: F ratio=10.87; Prob > F: =0.0021; iii) interaction paralog*taxonomy: F ratio=28.11; Prob > F: <0.0001.

Level	**Least Sq. Mean (COUSIN)**	**Standard error**	**Tukey's HSD group**
Paralog			
PTBP1	1.45	0.11	A
PTBP3	0.29	0.11	B
PTBP2	0.19	0.11	B
Taxonomy			
mammalian	0.44	0.080	A
non-mammalian	0.85	0.098	B
Paralog*Taxonomy			
PTBP1, mammalian	1.90	0.14	A
PTBP1, non-mammalian	0.99	0.17	B
PTBP2, non-mammalian	0.81	0.17	B
PTBP3, non-mammalian	0.75	0.17	B
PTBP3, mammalian	-0.16	0.14	C
PTBP2, mammalian	-0.43	0.14	C

Table 5: Comparison between species tree and subtrees of the nucleotide based maximum likelihood tree. Each subtree corresponds to a paralog. The K-tree score compares topological and pairwise distances between trees after re-scaling overall tree length, with higher values corresponding to more divergent trees. The Robinson-Foulds score compares only topological distances between trees, the values shown corresponding to the fraction of divergent nodes between trees.

Reference tree	**Comparison tree**	**K-tree score**	**Robinson-Foulds score**
Nucleotide tree VS species tree			
PTBP1	Species tree	0.759	42
PTBP2	Species tree	0.762	24
PTBP3	Species tree	1.700	28
Nucleotide tree VS Amino acid tree			
PTBP1-AA	*PTBP1*-NT	0.149	78
PTBP2-AA	*PTBP2*-NT	0.129	110
PTBP3-AA	*PTBP3*-NT	0.380	40

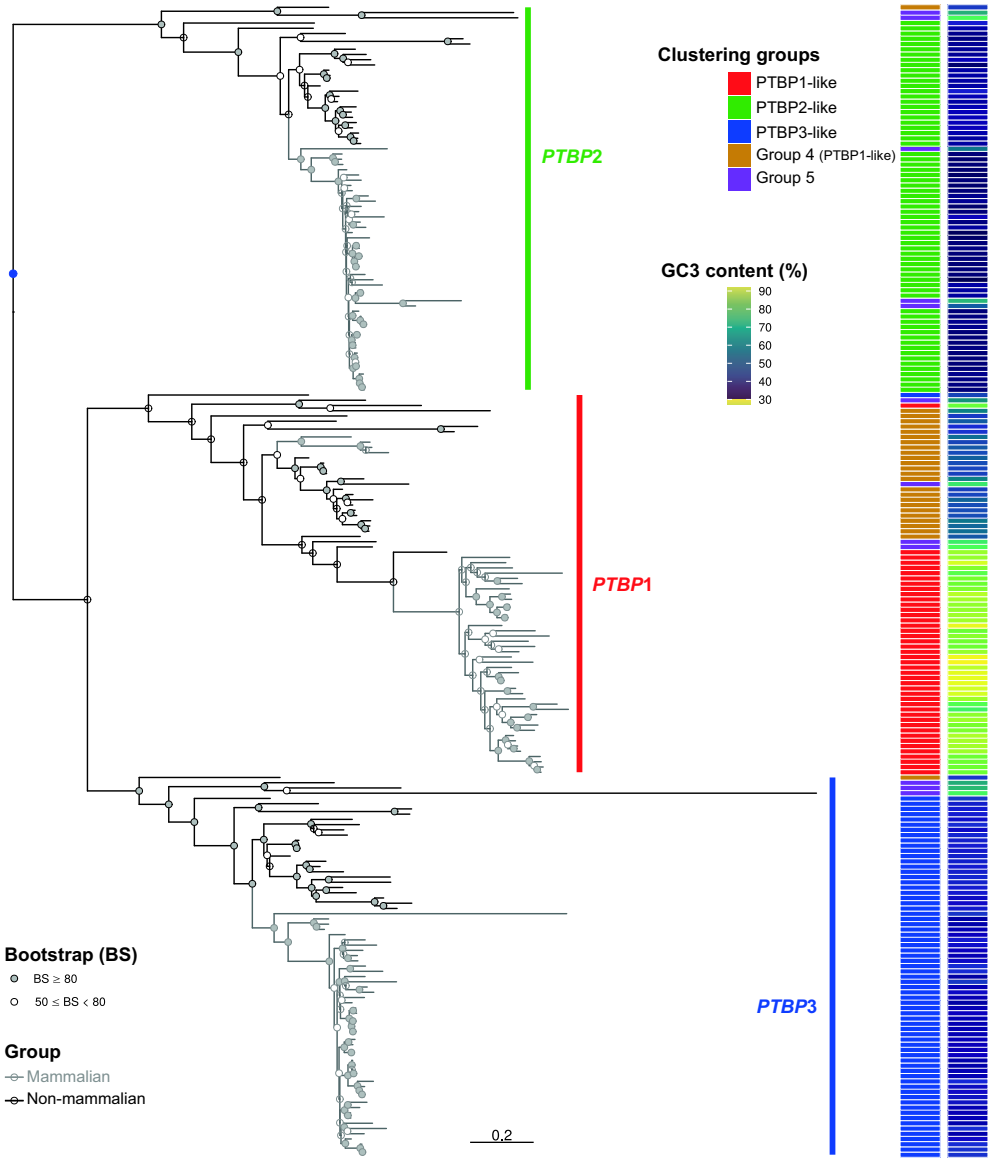

Figure 4: **Maximum-likelihood nucleic acid phylogeny of *PTBP*s genes.** The phylogram depicts *PTBP2*s (green side bar), *PTBP1*s (red side bar) and *PTBP3*s (blue side bar) clades. The outgroup genes from protostomata are not shown to focus on the scale for vertebrate *PTBP*s, but their placement on the tree and the polarity they provide for vertebrate *PTBP*s is given by the blue dot. Gray branches indicate mammalian *PTBP*s, while black branches indicate non-mammalian species. Note the lack of monophyly for mammals for *PTBP1*s. Filled dots on nodes indicate bootstrap values above 80, and empty dots indicate lower support values. Side bar on the left identifies the classification of each gene into the five groups identified by the hierarchical clusters, with the colour code in the inset. Side bar on the right displays GC3 content of the corresponding genes, with the gradient for the colour code ranging from 0 (blue) to 100% (yellow).

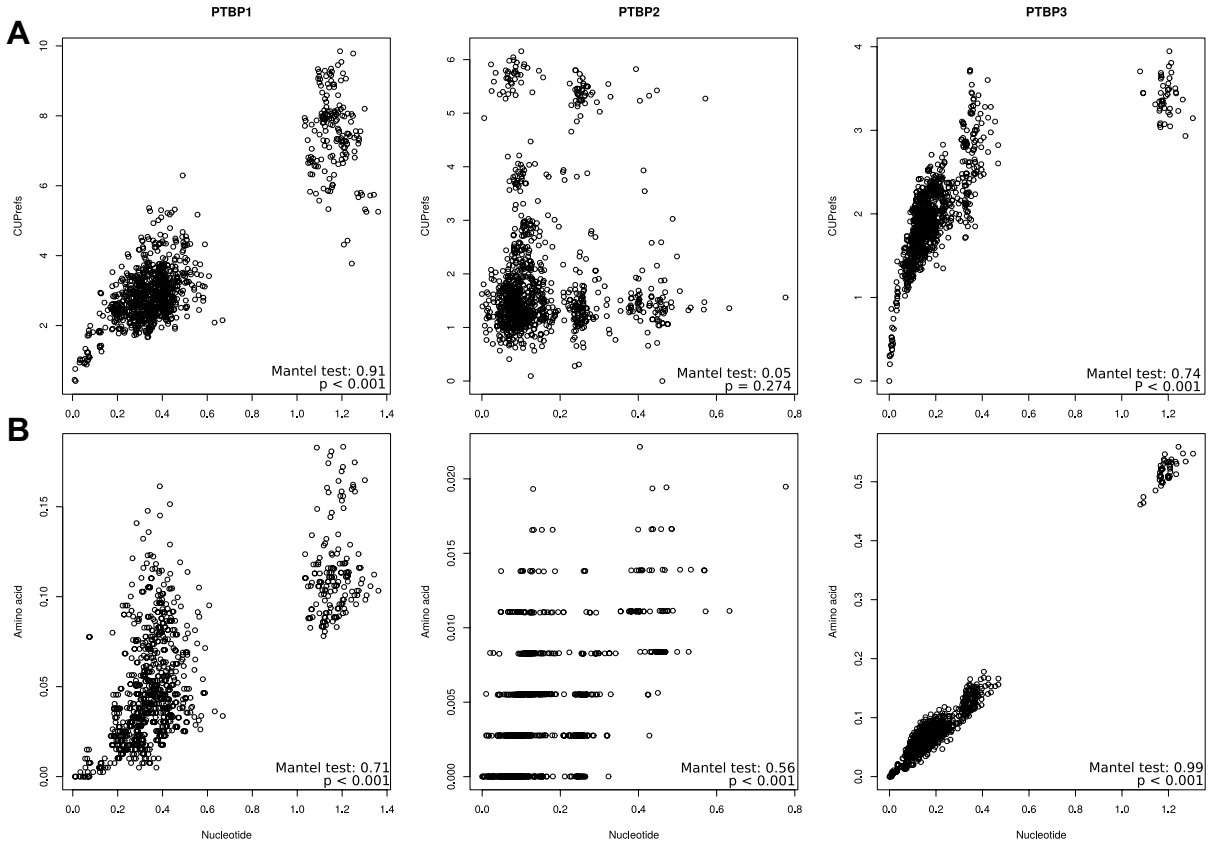

Figure 5: **Nucleotide-based pairwise distances against A) CUPrefs and B) amino-acid based pairwise distances for the different mammalian *PTBP* orthologs.** The results for a Mantel test assessing the correlation between the corresponding matrices are shown in the inset.

split apart from placental mammals in terms of both amino acid and nucleotide distances. This distribution matches well the fact that monotremes+marsupials do not cluster together with placental mammals in *PTBP1* phylogeny (see grey branches not being monophyletic for *PTBP1* in Figure 4). The same holds true for the platypus *PTBP3*, extremely divergent from the rest of the mammalian orthologs. For mammalian paralogs, the plots allow to see the increased number of overall mutations in general and of non-synonymous mutations in particular in *PTBP3*s compared with *PTBP1*. The precise mutational patterns are analysed in detail below. The histograms describing the accumulation of synonymous and non-synonymous mutations confirm that mammalian *PTBP1*s have selectively accumulated the largest number of synonymous mutations compared to non-mammalian *PTBP1*s and to other orthologs.

We have finally analysed the connection between nucleotide-based evolutionary distances within *PTBP* paralogs and CUPrefs-based distances (Figure 5A, Supplementary Material S8 A). A trend showing increased differences in CUPrefs as evolutionary distances increase is evident only for *PTBP1*s and *PTBP3*s in mammals. For mammalian *PTBP1*s the plot clearly differentiates a cloud with the values corresponding to the monotremes+marsupials splitting apart from placental mammals in terms of both evolutionary distance and CUPrefs. For mammalian *PTBP2*s the plot captures the divergent CUPrefs of the platypus and of the bats *M. natalensis* and *Desmodus rotundus*, while for non-

mammalian *PTBP2*s the divergent CUPrefs of the rainbow trout are obvious. Finally, for mammalian *PTBP3*s the large nucleotide divergence of the platypus paralog is evident. Importantly, all these instances of divergent behaviour (except for the platypus *PTBP3*) are consistent with the deviations described above from the expected composition by the mathematical modelling of the ortholog nucleotide composition.

Mammalian PTBP1s accumulate GC-enriching synonymous substitutions

We have shown that *PTBP1* genes are GC-richer and specifically GC3-richer than the *PTBP2* and *PTBP3* paralogs in the same genome, and that this enrichment is of a larger magnitude in placental *PTBP1*s. We have thus assessed whether a directional mutational pattern underlies this enrichment, especially regarding synonymous mutations. For this we have inferred the ancestral sequences of the respective most recent common ancestors of each *PTBP* paralogs, recapitulated synonymous and non-synonymous mutations between extant sequences and these ancestors, and constructed the corresponding mutation matrices (table S10). The two first axes of a principal component analysis using these mutational matrices capture, with a similar share, 66.95% of the variance between individuals (Figure 6). The first axis of the PCA separates synonymous from non-synonymous substitutions. Intriguingly though, while T<->C transitions are associated to synonymous mutations, as expected, G<->A transitions are associated to non-synonymous mutations. The second axis separates substitutions by their effect on nucleotide composition: GC-stabilizing/enriching on one direction, AT-stabilizing/enriching on the other one. Strikingly, the mutational spectrum of mammalian *PTBP1*s sharply differs from the rest of the paralogs. Substitutions in mammalian *PTBP1* towards GC-enriching changes, in both synonymous and non-synonymous compartments, are the main drivers of the second PCA axis. In contrast, synonymous mutations in *PTBP3* as well as all mutations in *PTBP2* tend to be AT-enriching. Finally, the mutational trends for *PTBP1* in mammals are radically different from those in non-mammals, while for *PTBP2* and *PTBP3*s the substitution patterns are similar in mammals and non-mammals for each of the compartments synonymous and non-synonymous.

4 Discussion

The non equal use of synonymous codons has puzzled biologists since first described. It has allowed for fruitful (and unfruitful) controversies between defenders of *all-is-neutralism* and defenders of *all-is-selectionism*, and has opened the door to the quest for embedded codes and signals behind CUPrefs patterns. The main questions around CUPrefs are twofold. On the one hand, their origin: to what extent they are the result of fine interplay between mutation and selection processes. On the other hand, their functional implications: whether and how particular CUPrefs can be linked to specific gene expression regulation processes, by modifying the kinetics and dynamics of DNA transcription, mRNA maturation and stability, mRNA translation, or protein folding and stability. In the present work we have built on the experimental results presented by Robinson and coworkers about the differential expression of the *PTBP* human gene paralogs as a function of their CUPrefs (Robinson et al., 2008). From this particular example, we have aimed at exploring by inductive thinking the general nature of the connection between paralogous gene evolution and CUPrefs. Our results show that the three *PTBP* paralogous genes of Vertebrates, which display divergent expression patterns, also have divergent nucleotide composition and CUPrefs. We propose here that this evolutionary pattern is compatible with a phenomenon of phenotypic evolution by sub-functionalisation (in this case specialisation in tissue-specific expression levels), associated to genotypic evolution by association to specific CUPrefs patterns.

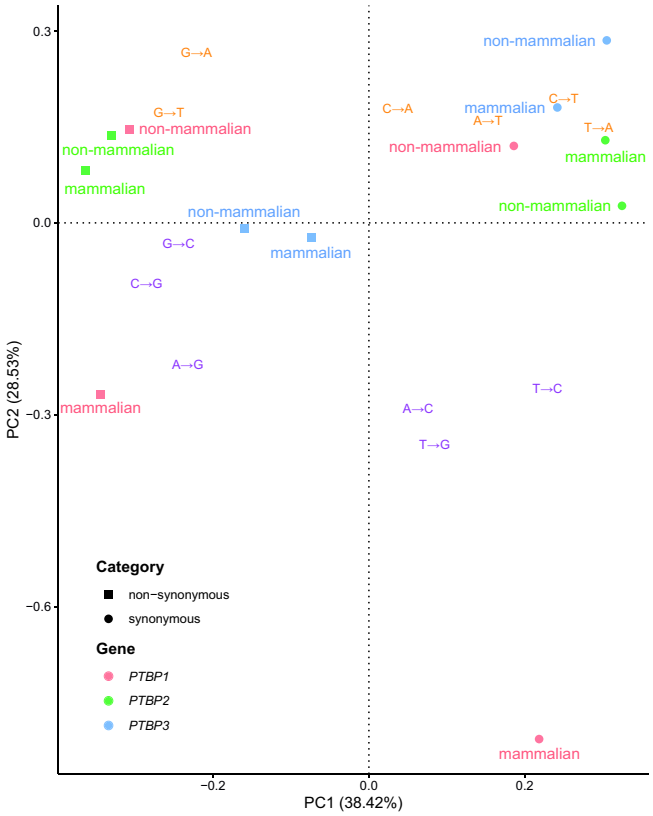

Figure 6: **Mutational spectra of synonymous and non-synonymous substitutions for *PTBP*s.** This principal component analysis (PCA) has been built using the observed nucleotide synonymous and non-synonymous substitution matrices for each *PTBP* paralog, inferred after phylogenetic inference and comparison of extant and ancestral sequences. The variables in this PCA are the types of substitution (*e.g.* A->G), identified by a colour code as GC-enriching / stabilizing substitutions (purple) or AT-enriching / stabilizing substitutions (orange). Variables are plotted according to their eigenvalues. Individuals in this PCA are the mutation categories in *PTBP* genes, stratified by their nature (synonymous or non-synonymous), by orthology (colour code for the different *PTBP*s is given in the inset) and by their taxonomy (mammals, or non-mammals).

We have reconstructed the phylogenetic relationships and analysed the evolution and diversity of CUPrefs among *PTBP* paralogs within 74 vertebrate species. The phylogenetic reconstruction shows that the genome of ancestral vertebrates already contained the three extant *PTBP* paralogs. This is consistent with the ortholog and paralog identification in the databases ENSEMBLE or ORTHOMAM (Yates et al., 2020; Scornavacca et al., 2019; Pina et al., 2018). Although our results suggest that *PTBP1* and *PTBP3* are sister lineages, the distant relationship of the vertebrate genes with the protostomate outgroup precludes the inference of a clear polarity between vertebrate *PTBP*s. We do not identify any instance of replacement between paralogs, and the evolutionary histories of the different *PTBP*s comply well with those of the corresponding species. The most blatant mismatch between gene and species trees is the polyphyly of mammalian *PTBP1*, with genes in monotremes and marsupials constituting a monophyletic clade, but not being the basal to and monophyletic with placental mammals. Multiple findings in our results point in this direction: i) the

excess of accumulation of synonymous mutations in mammalian *PTBP1*s for a similar total number of mutations (Figure 5 B); ii) the larger differences in CUPrefs between genes with a similar total number of nucleotide changes in the case of *PTBP1*s in mammals (Figure 5 A); iii) the explicitly different mutational spectrum of synonymous mutations in *PTBP1*s, enriched in A->C, T->G and T->C substitutions (Figure 6); iv) the sharp difference of CUPrefs between *PTBP1*s, and *PTBP2-3*s; and v) the clustering of *PTBP1* genes in monotremes and marsupials together with *PTBP1* genes in non-mammals according to their CUPrefs (Figure3 A). Overall, the particular nucleotide composition and the associated CUPrefs in mammalian *PTBP1* genes are most likely associated to specific mutational biases.

While GC3-rich nucleotide composition and CUPrefs of mammalian *PTBP1*s are dominated by local mutational biases, this is not the case for mammalian *PTBP2*s, overall AT3-richer. In vertebrates, nucleotide composition varies strongly along chromosomes, so that long stretches, historically named "isochores", appear enriched in GC or in AT nucleotides and present particular physico-chemical profiles (Caspersson et al., 1968). Local mutational biases underlying such heterogeneity, are the strongest evolutionary force shaping local nucleotide composition, so that the physical location of gene along the chromosome largely shapes its CUPrefs (Holmquist, 1989). In agreement with this mutational bias hypothesis, variation in GC3 composition of *PTBP1*s is almost totally explained by the variation in local GC composition (Table 3), suggesting that a same mutational bias has shaped the GC-rich composition of the flanking, intronic and coding regions of *PTBP1*s. The same trend, but to a lesser degree holds also true for *PTBP3*s. GC-biased gene conversion is often invoked as a powerful mechanism underlying such local GC-enrichment processes, leading to the systematic replacement of the alleles with the lowest GC composition by their GC richer homologs (Marais, 2003). It has been proposed that gene expression during meiosis facilitates GC-biased gene conversion during meiotic recombination (Pouyet et al., 2017), and in humans expression of *PTBP1*, GC3-enriched, is indeed documented during meiosis in the ovocite germinal line. Nevertheless, this line of reasoning does not holds true for *PTBP2*s. On the one hand, variations in local GC composition account barely for half of the variation in the *PTBP2* GC3 composition (Table 3). On the other hand, expression of *PTBP2*, AT3-enriched, is essential during spermatogenic meiosis (Zagore et al., 2015; Hannigan et al., 2017). Overall, GC3-enrichment in mammalian *PTBP2*s is compatible with GC-biased gene conversion events driving local mutational biases, but the AT3-enrichment of mammalian *PTBP2*s requires probably additional mechanisms to be explained, other than basal polymerase-related mutational biases for AT-enrichment, which acts as a background on the full genome (Hershberg and Petrov, 2010; Glémin et al., 2015; Petrov and Hartl, 1999).

In mammals, global GC-enriching genomic biases strongly impact CUPrefs, so that the most used codons in average tend to be GC-richer (Hershberg and Petrov, 2009). For this reason, in mammals GC3-rich *PTBP1*s match better the average genomic CUPrefs than AT3-richer *PTBP2*, which actually display CUPrefs in the opposite direction to the average of the genome. In the case of humans, *PTBP1* presents a COUSIN value of 1.747, consisting with an enrichment in preferentially-used codons, while on the contrary, the COUSIN value of -0.477 for *PTBP2* clearly points towards an enrichment in rare codons (Supplementary Material S4). Indeed, the poor match between human *PTBP2* CUPrefs and the human average CUPrefs results in poor expression of this gene in different human and murine cell lines, otherwise capable of expressing at high levels *PTBP1* and *PTBP3* (Robinson et al., 2008). The barrier to *PTBP2* expression seems to be the translation process, as *PTBP2* codon-recoding towards GC3-richer codons results in strong protein production in the same cellular context, without significant changes in the corresponding mRNA levels (Robinson et al., 2008). Such codon recoding strategy towards preferred codons has become indeed a standard practice

for gene expression engineering, despite our lack a comprehensive understanding of the impact and interaction on gene expression of local and global gene composition, nucleotide CUPrefs or mRNA structure (Brule and Grayhack, 2017).

The poor expression ability of human *PTBP2* in human cells, the large increase in protein production by the simple introduction of common codons and the lack of power of mutational biases to explain *PTBP2* nucleotide composition and CUPrefs, all raise the question of the adaptive value of the poor CUPrefs for this paralog. Specific tissue-dependent or cell-cycle dependent gene expression regulation patterns have been invoked to explain the codon usage-limited gene expression for certain human genes, such as *TLR7* or *KRAS* (Newman et al., 2016; Lampson et al., 2013; Fu et al., 2018). In humans, the expression levels of the three *PTBP* paralogs are tissue-dependent (Supplementary Material S1), and these differences are conserved through mammals (Keppetipola et al., 2012). In the case of the duplicated genes, subfunctionalisation through specialisation in spatio-temporal gene expression has often been proposed as the main evolutionary force driving conservation of paralogous genes (Ferris and Whitt, 1979). Such differential gene expression regulation in paralogs has actually been documented for a number of genes at very different taxonomic levels (Donizetti et al., 2009; Guschanski et al., 2017; Freilich et al., 2006). Specialised expression patterns in time and space can result in antagonistic presence/absence of the paralogous proteins (Adams et al., 2003). This is precisely the case of *PTBP1* and *PTBP2* during central nervous system development: in non-neuronal cells, *PTBP1* represses *PTBP2* expression by the skip of the exon 10 during *PTBP2* mRNA maturation, while during neuronal development, the micro RNA miR124 downregulates *PTBP1* expression, which in turn leads to upregulation of *PTBP2* (Keppetipola et al., 2012; Makeyev et al., 2007). Further, despite the high level of amino acid similarity between both proteins, *PTBP1* and *PTBP2* seem to perform complementary activities in the cell and to display different substrate specificity, so that they are not directly inter-exchangeable by exogenous manipulation of gene expression patterns (Vuong et al., 2016).

In a different subject, we want to drive the attention of the reader towards the puzzling trend of the UUG-Leu codon in our CUPrefs analyses. This UUG codon is the only GC-ending codon systematically clustering with AT-ending codons in all our analyses, and does not show the expected symmetrical behaviour with respect to UUA (see Figure 3). Such behaviour for UUG has been depicted, but not discussed, in other analyses of CUPrefs in mammalian genes (see figure 7 in Laurin-Lemay et al. (2018)), as well as for AGG-Arg and GGG-Gly in a global study of codon usages across the tree of life (see figure 1 in (Novoa et al., 2019)). The reasons underlying the clustering of UUG with AT-ending codons are unclear. A first line of thought could be functional: the UUG-Leu codon is particular because it can serve as alternative starting point for translation (Peabody, 1989). However, other codons such as ACG or GUG act more efficiently than UUG as translation initiation, and do not display any noticeable deviation (Ivanov et al., 2011). A second line of thought could be related to the tRNA repertoire, but both UUG and UUA are decoded by similar numbers of dedicated tRNAs in the vast majority of genomes (*e.g.* respectively six and seven tRNA genes in humans (Palidwor et al., 2010)). Finally, another line of thought suggests that UUG and AGG could be disfavoured if mutational pressure towards GC is very high, despite being GC-ending codons (Palidwor et al., 2010). Indeed, the series of synonymous transitions UUA->UUG->CUG for Leucine and the substitution chain AGA->AGG->CGG for Arginine are expected to lead to a depletion of UUG and of AGG codons when increasing GC content. Both UUG and ACG codons would this way display a non-linear, non-monotonic response to GC-mutational biases (Palidwor et al., 2010). In our dataset, however, AGG maps with the rest of GC-ending codons, symmetrically opposed to AGA as expected, and strongly contributing to the second PCA axis. Thus, only UUG presents frequency use patterns similar

to those of AT-ending codons. We humbly admit that we do not find a satisfactory explanation for this behaviour and invite researchers in the field to generate alternative explanatory hypotheses.

We have presented here an evolutionary analysis of the *PTBP* paralogs family, as a paradigm of evolution upon gene duplication. Our results show that CUPrefs in *PTBPs*s have evolved in parallel with specific gene expression regulation patterns. In the case of *PTBP1*, the most tissue-wise expressed of the paralogs, we have identified compositional, mutational biases as the driving force leading to strong enrichment in GC-ending codons. In contrast, for *PTBP2* the enrichment in AT-ending codons is rather compatible with selective forces related to specific spatio-temporal gene expression pattern, antagonistic to those of *PTBP1*. Our results suggest that the systematic study of composition, genomic location and expression patterns of paralogous genes can contribute to understanding the complex mutation-selection interplay shaping CUPrefs in multicellular organisms.

5 Acknowledgments

J.B. is the recipient of a PhD fellowship from the French Ministry of Education and Research. This study was supported by the European Union's Horizon 2020 research and innovation program under the grant agreement CODOVIREVOL (ERC-2014-CoG-647916) to I.G.B. The authors acknowledge the CNRS and the IRD for additional (meagre) intramural support. The computational results presented have been achieved in part using the IRD Bioinformatic Cluster itrop.

6 Data Availability Statement

All data required to reproduce our findings is provided in the tables in the main text or in the Supplementary Material section.

References

Adams KL, Cronn R, Percifield R, Wendel JF. 2003, April. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proceedings of the National Academy of Sciences of the United States of America. 100(8):4649–4654.

Apostolou-Karampelis K, Nikolaou C, Almirantis Y. 2016, August. A novel skew analysis reveals substitution asymmetries linked to genetic code GC-biases and PolIII a-subunit isoforms. DNA research: an international journal for rapid publication of reports on genes and genomes. 23(4):353–363.

Bourret J, Alizon S, Bravo IG. 2019, December. COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. Genome Biology and Evolution. 11(12):3523–3528. Publisher: Oxford Academic.

Brule CE, Grayhack EJ. 2017. Synonymous Codons: Choose Wisely for Expression. Trends in genetics: TIG. 33(4):283–297.

Bulmer M. 1991, November. The selection-mutation-drift theory of synonymous codon usage. Genetics. 129(3):897–907.

389 Caliskan N, Peske F, Rodnina MV. 2015, May. Changed in translation: mRNA recoding by 1 programmed ribosomal
390 frameshifting. Trends in Biochemical Sciences. 40(5):265–274.

391 Carbone A, Zinovyev A, Képès F. 2003, November. Codon adaptation index as a measure of dominating codon bias.
392 Bioinformatics (Oxford, England). 19(16):2005–2015.

393 Caspersson T, Farber S, Foley GE, Kudynowski J, Modest EJ, Simonsson E, Wagh U, Zech L. 1968, January. Chemical
394 differentiation along metaphase chromosomes. Experimental Cell Research. 49(1):219–222.

395 Castresana J. 2000, April. Selection of conserved blocks from multiple alignments for their use in phylogenetic
396 analysis. Molecular Biology and Evolution. 17(4):540–552.

397 Chamary JV, Parmley JL, Hurst LD. 2006, February. Hearing silence: non-neutral evolution at synonymous sites in
398 mammals. Nature Reviews. Genetics. 7(2):98–108.

399 Clark JM. 1988, October. Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic
400 DNA polymerases. Nucleic Acids Research. 16(20):9677–9686.

401 Copley SD. 2020, April. Evolution of new enzymes by gene duplication and divergence. The FEBS journal.
402 287(7):1262–1283.

403 Donizetti A, Fiengo M, Minucci S, Aniello F. 2009, October. Duplicated zebrafish relaxin-3 gene shows a different
404 expression pattern from that of the co-orthologue gene. Development, Growth & Differentiation. 51(8):715–722.

405 Duret L. 2002, December. Evolution of synonymous codon usage in metazoans. Current Opinion in Genetics &
406 Development. 12(6):640–649.

407 Duret L, Mouchiroud D. 1999, April. Expression pattern and, surprisingly, gene length shape codon usage in
408 Caenorhabditis, Drosophila, and Arabidopsis. Proceedings of the National Academy of Sciences. 96(8):4482–4487.
409 Publisher: National Academy of Sciences Section: Biological Sciences.

410 Ferris SD, Whitt GS. 1979, April. Evolution of the differential regulation of duplicate genes after polyploidization.
411 Journal of Molecular Evolution. 12(4):267–317.

412 Freilich S, Massingham T, Blanc E, Goldovsky L, Thornton JM. 2006. Relating tissue specialization to the differenti-
413 ation of expression of singleton and duplicate mouse proteins. Genome Biology. 7(10):R89.

414 Fu J, Dang Y, Counter C, Liu Y. 2018. Codon usage regulates human KRAS expression at both transcriptional and
415 translational levels. The Journal of Biological Chemistry. 293(46):17929–17940.

416 Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. 2018, May. Codon Usage
417 Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene
418 Conversion. Molecular Biology and Evolution. 35(5):1092–1103.

419 Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015, August. Quantification of GC-biased gene
420 conversion in the human genome. Genome Research. 25(8):1215–1228. Company: Cold Spring Harbor Laboratory
421 Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label:
422 Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

423 Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980, January. Codon catalog usage and the genome hypothesis.
424 Nucleic Acids Research. 8(1):r49–r62.

425 Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs.
426 Genome Research. 27(9):1461–1474.

427 Hannigan MM, Zagore LL, Licatalosi DD. 2017, June. Ptbp2 controls an alternative splicing network required for cell
428 communication during spermatogenesis. Cell reports. 19(12):2598–2612.

429 Hershberg R, Petrov DA. 2009, July. General rules for optimal codon choice. PLoS genetics. 5(7):e1000556.

430 Hershberg R, Petrov DA. 2010, September. Evidence That Mutation Is Universally Biased towards AT in Bacteria.
431 PLoS Genetics. 6(9).

432 Holmquist GP. 1989, June. Evolution of chromosome bands: Molecular ecology of noncoding DNA. Journal of
433 Molecular Evolution. 28(6):469–486.

434 Ikemura T. 1981, September. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence
435 of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E.
436 coli translational system. Journal of Molecular Biology. 151(3):389–409.

437 Ivanov IP, Firth AE, Michel AM, Atkins JF, Baranov PV. 2011, May. Identification of evolutionarily conserved non-
438 AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Research. 39(10):4220–4234.

439 Katoh K, Misawa K, Kuma Ki, Miyata T. 2002, July. MAFFT: a novel method for rapid multiple sequence alignment
440 based on fast Fourier transform. Nucleic Acids Research. 30(14):3059–3066.

441 Keppetipola N, Sharma S, Li Q, Black DL. 2012, August. Neuronal regulation of pre-mRNA splicing by polypyrim-
442 idine tract binding proteins, PTBP1 and PTBP2. Critical Reviews in Biochemistry and Molecular Biology.
443 47(4):360–378.

444 Khorana HG, Büchi H, Ghosh H, Gupta N, Jacob TM, Kössel H, Morgan R, Narang SA, Ohtsuka E, Wells RD. 1966.
445 Polynucleotide synthesis and the genetic code. Cold Spring Harbor Symposia on Quantitative Biology. 31:39–49.

446 Koonin EV. 2005. Orthologs, Paralogs, and Evolutionary Genomics. Annual Review of Genetics. 39(1):309–338.
447 _eprint: https://doi.org/10.1146/annurev.genet.39.073003.114725.

448 Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence
449 Times. Molecular Biology and Evolution. 34(7):1812–1819.

450 Lampson BL, Pershing NLK, Prinz JA, Lacsina JR, Marzluff WF, Nicchitta CV, MacAlpine DM, Counter CM. 2013,
451 January. Rare codons regulate KRas oncogenesis. Current biology: CB. 23(1):70–75.

452 Laurin-Lemay S, Rodrigue N, Lartillot N, Philippe H. 2018. Conditional Approximate Bayesian Computation: A New
453 Approach for Across-Site Dependency in High-Dimensional Mutation-Selection Models. Molecular Biology and
454 Evolution. 35(11):2819–2834.

455 Lujan SA, Williams JS, Pursell ZF, Abdulovic-Cui AA, Clark AB, McElhinny SAN, Kunkel TA. 2012, October. Mis-
456 match Repair Balances Leading and Lagging Strand DNA Replication Fidelity. PLOS Genetics. 8(10):e1003016.
457 Publisher: Public Library of Science.

458 Makeyev EV, Zhang J, Carrasco MA, Maniatis T. 2007, August. The MicroRNA miR-124 Promotes Neuronal Differ-
459 entiation by Triggering Brain-Specific Alternative Pre-mRNA Splicing. Molecular cell. 27(3):435–448.

460 Marais G. 2003, June. Biased gene conversion: implications for genome and sex evolution. Trends in Genetics.
461 19(6):330–338. Publisher: Elsevier.

462 Mordstein C, Savisaar R, Young RS, Bazile J, Talmane L, Luft J, Liss M, Taylor MS, Hurst LD, Kudla G. 2020, April.
463 Codon Usage and Splicing Jointly Influence mRNA Localization. Cell Systems. 10(4):351–362.e8.

464 NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. Nu-
465 cleic Acids Research. 46(D1):D8–D13.

466 Newman ZR, Young JM, Ingolia NT, Barton GM. 2016, March. Differences in codon bias and GC content contribute
467 to the balanced expression of TLR7 and TLR9. Proceedings of the National Academy of Sciences of the United
468 States of America. 113(10):E1362–1371.

469 Nirenberg MW, Matthaei JH. 1961, October. THE DEPENDENCE OF CELL- FREE PROTEIN SYNTHESIS IN E.
470 COLI UPON NATURALLY OCCURRING OR SYNTHETIC POLYRIBONUCLEOTIDES. Proceedings of the
471 National Academy of Sciences of the United States of America. 47(10):1588–1602.

472 Novoa EM, Jungreis I, Jaillon O, Kellis M. 2019. Elucidation of Codon Usage Signatures across the Domains of Life.
473 Molecular Biology and Evolution. 36(10):2328–2339.

474 Novoa EM, Ribas de Pouplana L. 2012, November. Speeding with control: codon usage, tRNAs, and ribosomes.
475 Trends in genetics: TIG. 28(11):574–581.

476 Palidwor GA, Perkins TJ, Xia X. 2010, October. A general model of codon bias due to GC mutational bias. PloS One.
477 5(10):e13431.

478 Peabody DS. 1989, March. Translation initiation at non-AUG triplets in mammalian cells. The Journal of Biological
479 Chemistry. 264(9):5031–5035.

480 Percudani R, Pavesi A, Ottonello S. 1997, May. Transfer RNA gene redundancy and translational selection in Saccha-
481 romyces cerevisiae11Edited by J. Karn. Journal of Molecular Biology. 268(2):322–330.

482 Petrov DA, Hartl DL. 1999, February. Patterns of nucleotide substitution in Drosophila and mammalian genomes.
483 Proceedings of the National Academy of Sciences. 96(4):1475–1479. Publisher: National Academy of Sciences
484 Section: Biological Sciences.

485 Pina J, Ontiveros RJ, Keppetipola N, Nikolaidis N. 2018, April. A Bioinformatics Approach to Discover the Evolu-
486 tionary Origin of the PTBP Splicing Regulators. The FASEB Journal. 32(1_supplement):802.16–802.16. Publisher:
487 Federation of American Societies for Experimental Biology.

488 Plotkin JB, Kudla G. 2011, January. Synonymous but not the same: the causes and consequences of codon bias. Nature
489 Reviews Genetics. 12(1):32–42.

490 Pouyet F, Mouchiroud D, Duret L, Sémon M. 2017. Recombination, meiotic expression and human codon usage.
491 eLife. 6.

492 Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR,
493 Coller J. 2015, March. Codon optimality is a major determinant of mRNA stability. Cell. 160(6):1111–1124.

494 Reijns MAM, Kemp H, Ding J, Marion de Procé S, Jackson AP, Taylor MS. 2015, February. Lagging-strand replication
495 shapes the mutational landscape of the genome. Nature. 518(7540):502–506. Number: 7540 Publisher: Nature
496 Publishing Group.

497 Robinson DF, Foulds LR. 1981, February. Comparison of phylogenetic trees. Mathematical Biosciences. 53(1):131–
498 147.

499 Robinson F, Jackson RJ, Smith CWJ. 2008, March. Expression of Human nPTB Is Limited by Extreme Suboptimal
500 Codon Content. PLOS ONE. 3(3):e1801. Publisher: Public Library of Science.

501 Satapathy SS, Powdel BR, Buragohain AK, Ray SK. 2016, October. Discrepancy among the synonymous codons
502 with respect to their selection as optimal codon in bacteria. DNA Research. 23(5):441–449. Publisher: Oxford
503 Academic.

504 Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019, April. OrthoMaM v10:
505 Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian
506 Genomes. Molecular Biology and Evolution. 36(4):861–862. Publisher: Oxford Academic.

507 Sharp PM, Li WH. 1987. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and
508 its potential applications. Nucleic Acids Research. 15(3):1281–1295.

509 Sonnhammer ELL, Koonin EV. 2002, December. Orthology, paralogy and proposed classification for paralog subtypes.
510 Trends in genetics: TIG. 18(12):619–620.

511 Soria-Carrasco V, Talavera G, Igea J, Castresana J. 2007, November. The K tree score: quantification of differences
512 in the relative branch length and topology of phylogenetic trees. Bioinformatics (Oxford, England). 23(21):2954–
513 2956.

514 Spencer PS, Barral JM. 2012, March. Genetic code redundancy and its influence on the encoded polypeptides. Com-
515 putational and Structural Biotechnology Journal. 1.

516 Stamatakis A. 2014, May. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.
517 Bioinformatics (Oxford, England). 30(9):1312–1313.

518 Vuong JK, Lin CH, Zhang M, Chen L, Black DL, Zheng S. 2016. PTBP1 and PTBP2 Serve Both Specific and
519 Redundant Functions in Neuronal Pre-mRNA Splicing. Cell Reports. 17(10):2766–2775.

520 Whittle CA, Extavour CG. 2016, September. Expression-Linked Patterns of Codon Usage, Amino Acid Frequency,
521 and Protein Length in the Basally Branching Arthropod Parasteatoda tepidariorum. Genome Biology and Evolution.
522 8(9):2722–2736. Publisher: Oxford Academic.

523 Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett
524 R, Bhai J, Billis K, Boddu S, Marugán JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil
525 L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T,
526 Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker
527 A, Parton A, Patricio M, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M,
528 Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M,
529 Flint B, Frankish A, Hunt SE, IIsley G, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge
530 JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR, Flicek P. 2020, January.
531 Ensembl 2020. Nucleic Acids Research. 48(D1):D682–D688. Publisher: Oxford Academic.

532 Zagore LL, Grabinski SE, Sweet TJ, Hannigan MM, Sramkoski RM, Li Q, Licatalosi DD. 2015, December. RNA
533 Binding Protein Ptbp2 Is Essential for Male Germ Cell Development. Molecular and Cellular Biology. 35(23):4030–
534 4042.

B

Annexe Deux

ANNEXES CHAPITRE UN

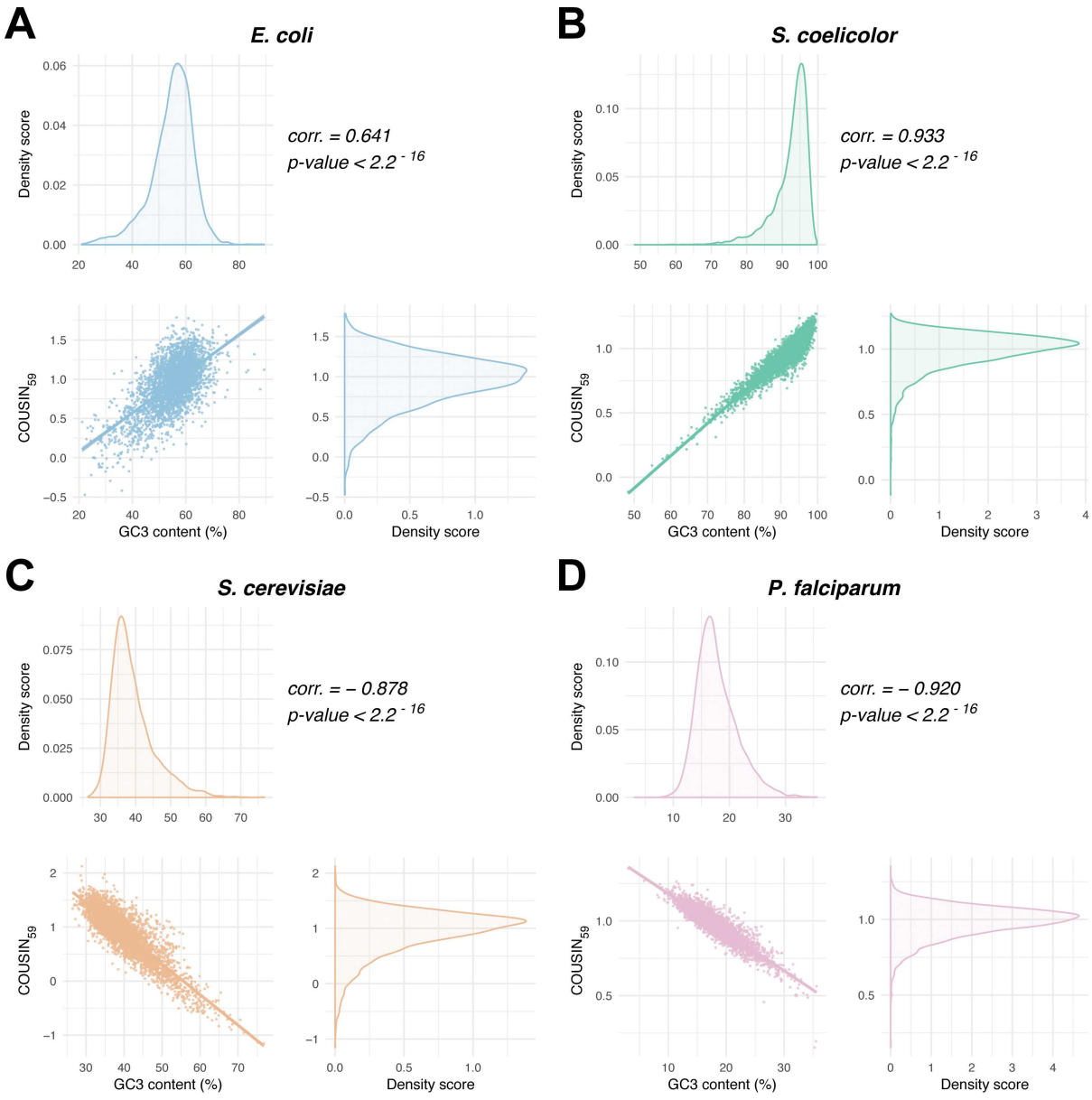

Figure B.1 – Scatterplot of COUSIN_{59} (x-axis) and GC3 content (y-axis) scores for CDSs belonging to *E. coli* (A), *S. coelicolor* (B), *S. cerevisiae* (C) and *P. falciparum* (D). Each scatterplot is accompanied by two density curves : COUSIN_{59} (right of the scatterplot) and GC3 content (top of the scatterplot). On the top-right of the scatterplots, statistics of Pearson correlation tests between COUSIN_{59} scores and GC3 content is given.

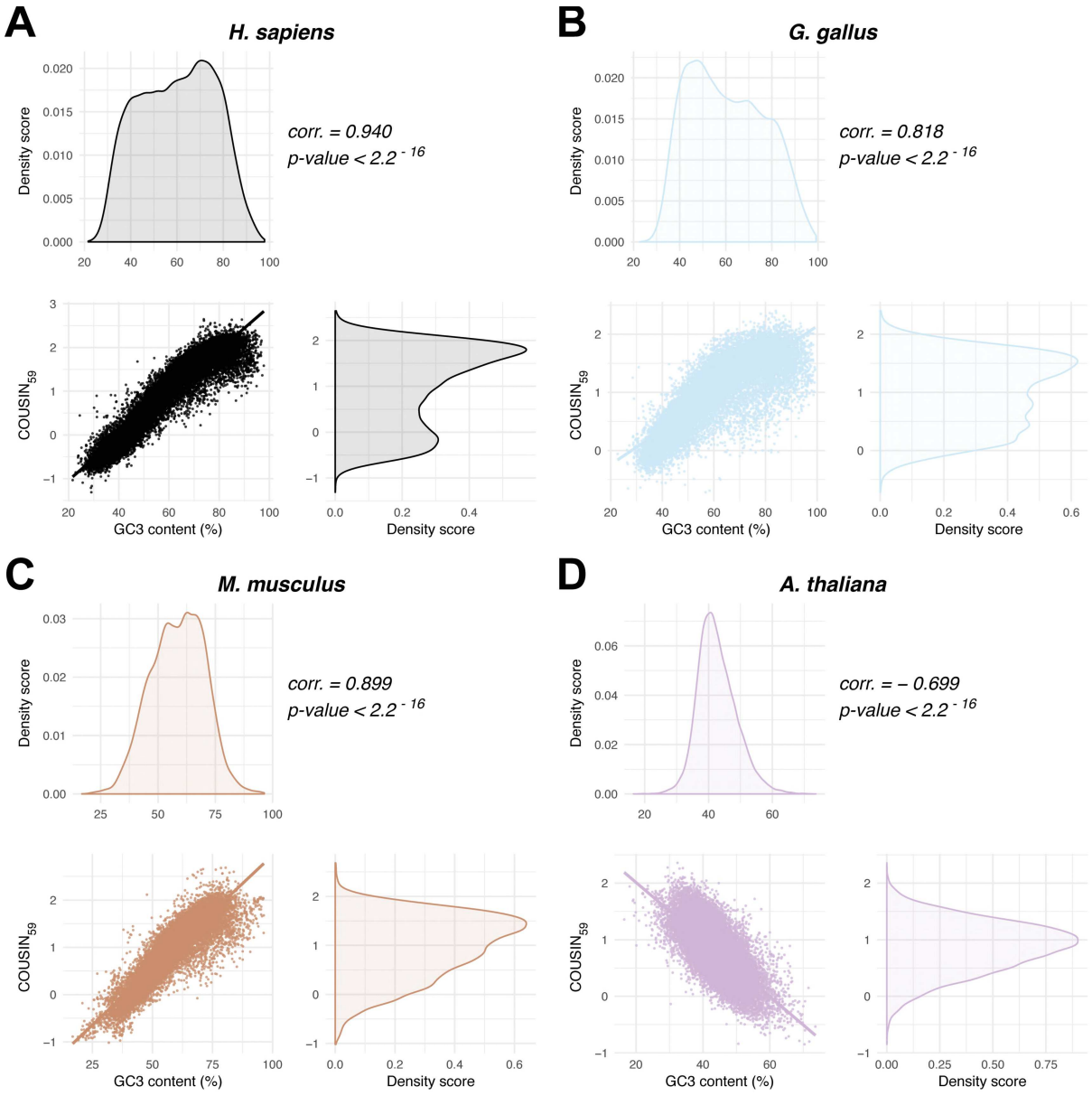

Figure B.2 – Scatterplot of COUSIN_{59} (x-axis) and GC3 content (y-axis) scores for CDSs belonging to *H. sapiens* (A), *G. gallus* (B), *Mus musculus* (C) and *A. thaliana* (D). Each scatterplot is accompanied by two density curves : COUSIN_{59} (right of the scatterplot) and GC3 content (top of the scatterplot). On the top-right of the scatterplots, statistics of Pearson correlation tests between COUSIN_{59} scores and GC3 content is given.

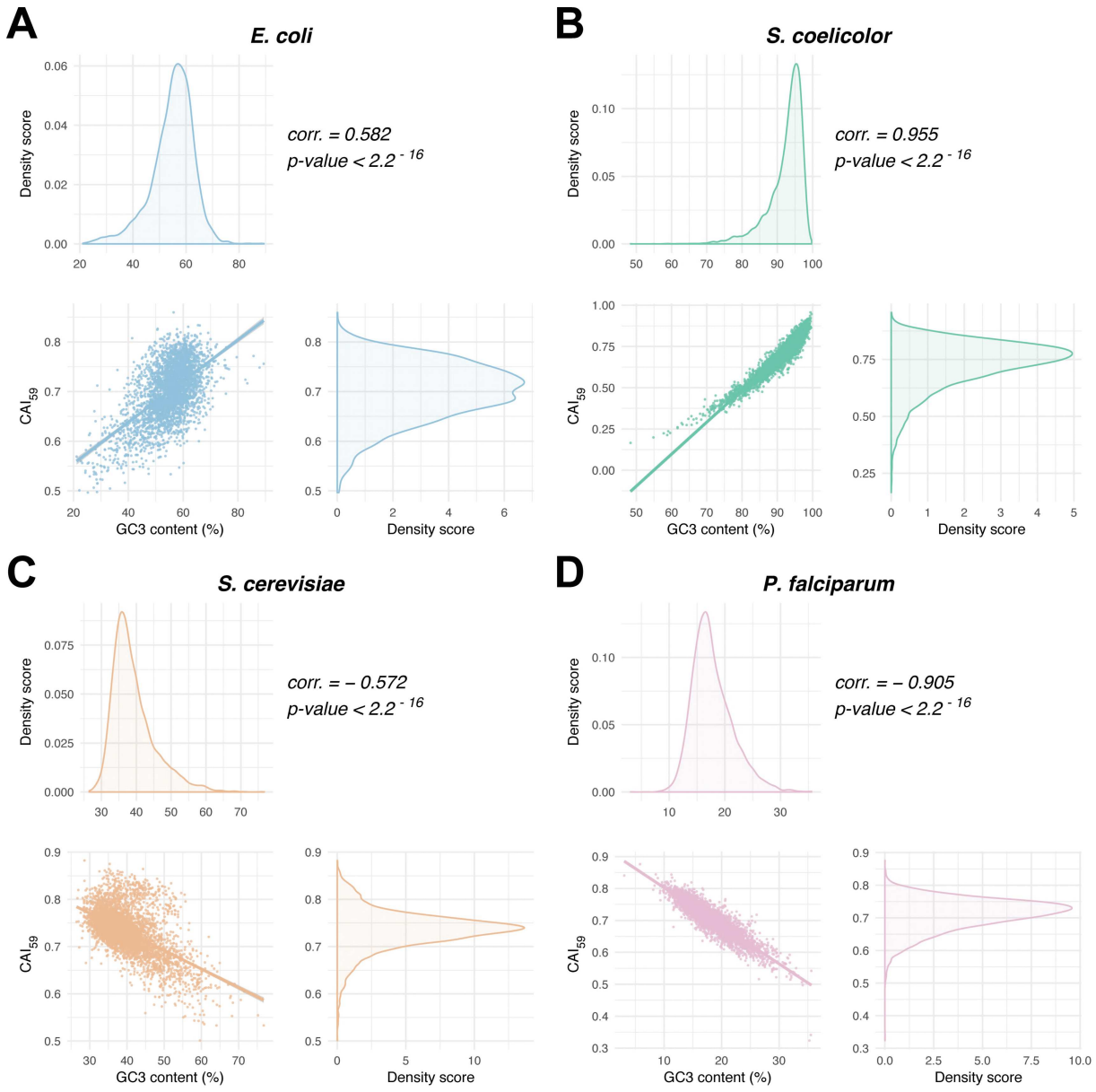

Figure B.3 – Scatterplot of CAI_{59} (x-axis) and GC3 content (y-axis) scores for CDSs belonging to *E. coli* (A), *S. coelicolor* (B), *S. cerevisiae* (C) and *P. falciparum* (D). Each scatterplot is accompanied by two density curves : CAI_{59} (right of the scatterplot) and GC3 content (top of the scatterplot). On the top-right of the scatterplots, statistics of Pearson correlation tests between CAI_{59} scores and GC3 content is given.

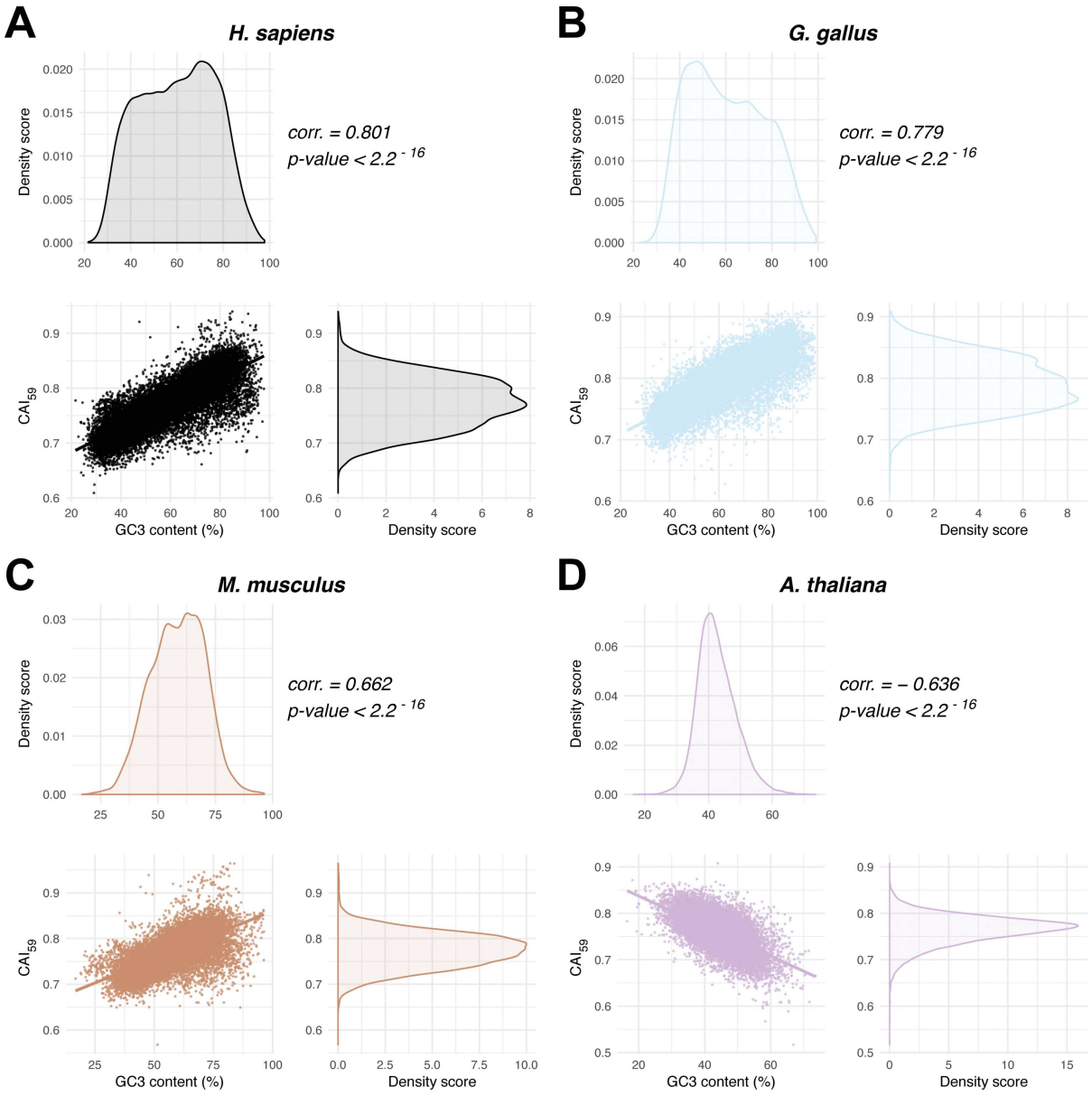

Figure B.4 – Scatterplot of CAI_{59} (x-axis) and GC3 content (y-axis) scores for CDSs belonging to *H. sapiens* (A), *G. gallus* (B), *Mus musculus* (C) and *A. thaliana* (D). Each scatterplot is accompanied by two density curves : CAI_{59} (right of the scatterplot) and GC3 content (top of the scatterplot). On the top-right of the scatterplots, statistics of Pearson correlation tests between CAI_{59} scores and GC3 content of CDSs is given.

TABLE B.1 – **Moyenne du % de GC3 des CDS complets pour chacun des organismes de cette étude.**. Cette table décrit les contenu en GC3 à l'échelle de tous les CDS, mais aussi des CDS ayant les scores $COUSIN_{59}$ les plus hauts, les plus bas, et le reste des CDS ne rentrant pas dans ces deux catégories.

	E. coli	***S. coelicolor***	***S. cerevisiae***	***P. falciparum***	***H. sapiens***	***G. gallus***	***M. musculus***	***A. thaliana***
contenu en GC3 (jeu complet)	54.9	92.3	39.2	17.8	59.8	60.6	58.6	42.7
contenu en GC3 (haut 20%)	60.0	96.6	34.0	13.9	79.0	77.9	71.8	37.8
contenu en GC3 (milieu 60%)	56.1	93.5	38.0	17.3	61.0	61.2	59.5	42.1
contenu en GC3 (bas 20%)	46.1	84.7	48.1	23.1	37.7	41.6	42.8	49.4

TABLE B.2 – Mean value, Huber-M estimator value and MAD scores of $COUSIN_{18}$, $COUSIN_{59}$, CAI_{18} and CAI_{59} scores on the studied organisms.

	E. coli	***S. coelicolor***	***S. cerevisiae***	***P. falciparum***	***H. sapiens***	***G. gallus***	***M. musculus***	***A. thaliana***
Mean ($COUSIN_{18}$)	0.93	0.98	0.91	0.98	0.98	0.99	0.98	0.85
Huber-M estimator ($COUSIN_{18}$)	0.94	0.99	0.95	0.98	0.98	0.99	0.99	0.86
MAD ($COUSIN_{18}$)(+/-)	0.33	0.10	0.33	0.10	1.23	0.85	0.81	0.50
Mean ($COUSIN_{59}$)	0.94	0.98	0.93	0.98	0.95	0.97	0.97	0.87
Huber-M estimator ($COUSIN_{59}$)	0.96	1.00	0.97	0.99	0.95	0.97	0.98	0.88
MAD ($COUSIN_{59}$)(+/-)	0.29	0.11	0.32	0.00	1.03	0.74	0.67	0.45
Mean (CAI_{18})	0.74	0.77	0.74	0.71	0.81	0.82	0.81	0.79
Huber-M estimator (CAI_{18})	0.74	0.78	0.74	0.71	0.81	0.23	0.81	0.79
MAD (CAI_{18})(+/-)	0.05	0.09	0.03	0.05	0.05	0.05	0.04	0.02
Mean (CAI_{59})	0.70	0.73	0.73	0.71	0.77	0.79	0.77	0.76
Huber-M estimator (CAI_{59})	0.70	0.74	0.74	0.71	0.77	0.79	0.77	0.77
MAD (CAI_{59})(+/-)	0.06	0.09	0.03	0.04	0.05	0.05	0.04	0.03

TABLE B.3 – Size, number of CDSs, Huber-M estimator values and MAD values for GC3 and $COUSIN_{59}$ among *G. gallus* chromosomes

Chromosome	Size (Mb)	CDSs	Huber-M estimator (GC3)	MAD (+/-)(GC3)	Huber-M estimator ($COUSIN_{59}$)	MAD (+/-) ($COUSIN_{59}$)
1	197.6	2025	53.8	14.2	0.8	0.7
2	149.7	1315	51.4	12.5	0.7	0.6
3	110.8	1124	53.1	14.0	0.7	0.7
4	91.3	1094	56.0	16.4	0.8	0.7
5	59.8	920	58.3	17.6	0.9	0.8
6	36.4	520	57.5	16.9	0.9	0.7
7	36.7	470	54.3	15.3	0.8	0.7
8	30.2	491	57.1	19.1	0.9	0.8
9	24.2	415	59.8	19.0	0.9	0.8
10	21.1	400	60.4	19.9	1.0	0.8
11	20.2	356	63.5	22.8	1.1	0.7
12	20.3	340	64.0	22.4	1.0	0.7
13	19.1	354	65.7	20.2	1.1	0.6
14	16.2	392	63.8	18.7	1.1	0.6
15	13.1	347	63.6	18.1	1.1	0.6
16	2.8	133	70.3	10.5	1.3	0.37
17	10.8	284	66.2	16.5	1.2	0.6
18	11.4	306	66.3	19.0	1.2	0.6
19	10.3	324	65.7	16.8	1.2	0.6
20	13.9	335	64.4	18.2	1.2	0.6
21	6.8	236	65.0	17.2	1.2	0.5
22	5.5	186	73.8	13.4	1.4	0.4
23	6.2	248	69.9	17.4	1.3	0.5
24	6.5	184	68.1	15.0	1.4	0.4
25	4.0	259	76.4	12.2	1.5	0.4
26	6.1	268	70.7	14.3	1.4	0.4
27	8.1	296	75.1	13.1	1.5	0.4
28	5.1	308	73.1	14.5	1.4	0.4
30	1.8	79	79.9	6.4	1.3	0.4
31	6.2	213	67.8	4.4	1.3	0.4
32	0.7	55	84.6	5.7	1.4	0.3
33	7.8	463	73.7	11.8	1.4	0.3
W	6.8	37	46.6	13.3	0.4	0.5
Z	82.5	773	52.6	15.8	0.7	0.7

C

Annexe Trois

ANNEXES CHAPITRE DEUX

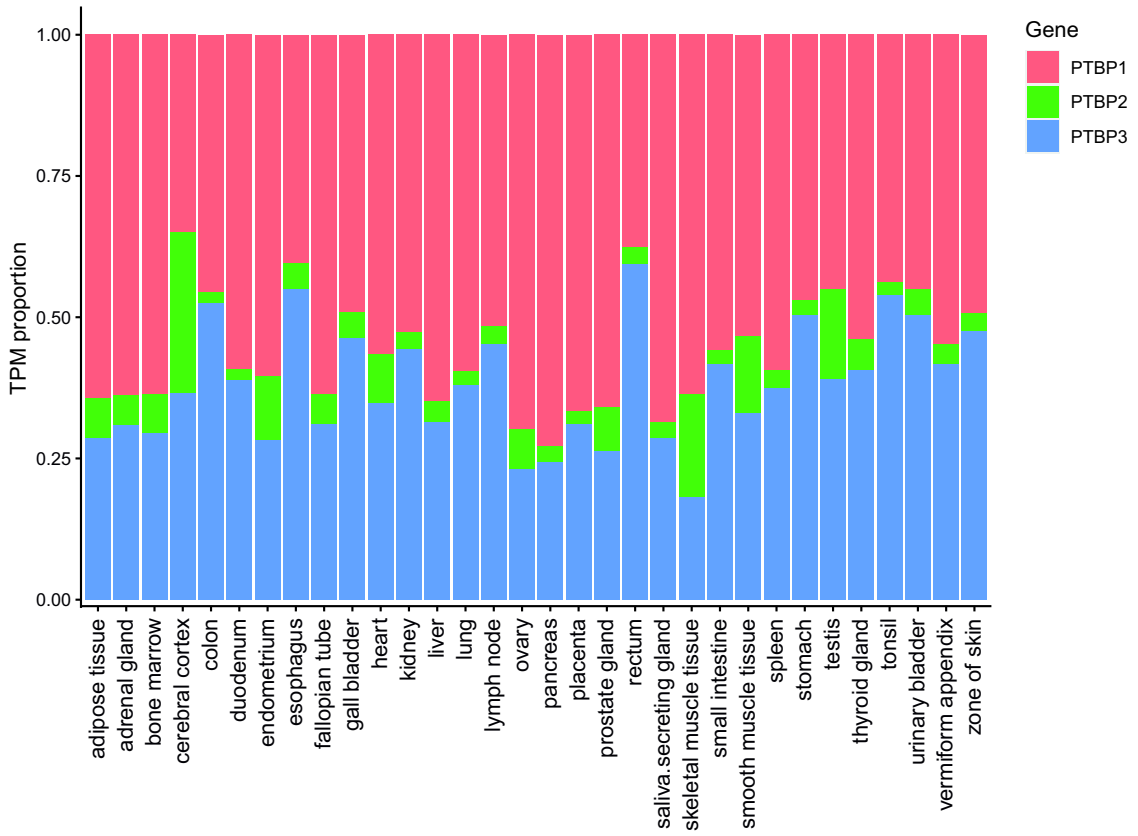

Figure C.1 – **Relative mRNA levels of *PTBP1* (red), *PTBP2* (green) and *PTBP3* (blue) in 32 human tissues.** Values given in this graph represent the proportion of transcripts per million (TPM) for each *PTBP*, as communicated by Zhang and coworkers [141] . Overall *PTBP1* is the most transcribed gene followed by *PTBP3*, while *PTBP2* shows maximum expression in cerebral cortex and in skeletal muscle.

TABLE C.1 – **Percentages of total GC content and GC content at the first (GC1), second (GC2) and third (GC3) position of nucleotides for *PTBP1*, *PTBP2* and *PTBP3***. Lines in bold present the mean score values for mammalians and non-mammalians species.

	PTBP1				***PTBP2***				***PTBP3***			
Species	**%C**	**%C1**	**%C2**	**%C3**	**%C**	**%C1**	**%C2**	**%C3**	**%C**	**%C1**	**%C2**	**%C3**
Mammalians	**59.71**	**58.16**	**41.19**	**79.77**	**41.88**	**55.72**	**38.25**	**31.66**	**44.51**	**56.30**	**41.01**	**36.23**
Aotus nancymaae	61.33	58.29	40.76	84.95	41.20	55.56	38.29	29.76	44.25	55.94	41.57	35.25
Balaenoptera acutorostrata	61.65	58.10	41.14	85.71	41.20	55.95	38.29	29.37	43.81	56.90	41.00	33.53
Bos taurus	59.30	58.29	41.52	78.10	41.53	55.75	38.29	30.56	44.87	56.79	40.92	36.90
Canis lupus	60.00	58.48	40.95	80.57	41.53	55.75	38.29	30.56	44.49	56.57	41.10	35.81
Cebus capucinus	61.19	57.92	39.08	86.57	41.47	55.56	38.29	30.56	44.19	56.13	41.38	35.06
Ceratotherium simum	62.35	58.67	41.33	87.05	41.07	55.36	38.29	29.56	44.57	56.46	40.66	36.61
Chinchilla lanigera	58.92	57.52	41.33	77.91	42.06	55.75	38.29	32.14	44.72	56.43	40.88	36.85
Chrysochloris asiatica	58.86	58.29	41.33	76.95	40.62	55.31	38.08	28.46	43.97	55.39	40.96	35.58
Condylura cristata	61.08	57.71	41.33	84.19	41.34	55.95	38.29	29.76	43.31	55.47	40.69	33.78
Dasypus novemcinctus	65.08	61.17	43.41	90.66	40.87	55.36	38.29	28.97	44.06	55.75	41.00	35.44
Delphinapterus leucas	61.33	57.71	41.33	84.95	41.15	55.51	38.08	29.86	44.19	57.28	41.19	34.10
Desmodus rotundus	56.44	57.91	40.95	70.48	51.98	57.94	38.49	59.52	43.61	55.75	40.81	34.29
Echinops telfairi	60.96	59.89	41.07	81.94	40.99	55.29	37.92	29.74	44.96	56.51	41.00	37.36
Equus asinus	63.55	58.59	41.22	90.84	41.40	55.56	38.29	30.36	45.86	56.84	40.46	40.27
Equus caballus	63.61	58.78	41.22	90.84	41.67	55.56	38.29	31.15	46.05	56.84	40.46	40.85
Equus przewalskii	63.15	57.39	41.46	90.60	41.67	55.56	38.29	31.15	46.05	56.84	40.46	40.85
Erinaceus europaeus	63.80	58.78	40.84	91.79	40.34	55.75	37.90	27.38	43.18	54.58	41.34	33.61
Felis catus	60.64	58.48	40.95	82.48	41.27	55.75	38.29	29.76	44.36	56.15	40.96	35.96
Heterocephalus glaber	59.75	57.91	41.52	79.81	41.47	55.75	38.29	30.36	43.81	56.32	40.81	34.29
Homo sapiens	61.27	58.67	40.76	84.38	41.07	55.56	38.29	29.37	43.93	56.07	41.23	34.49
Ictidomys tridecemlineatus	60.06	57.52	41.33	81.33	41.20	55.75	38.29	29.56	43.17	55.36	41.00	33.14
Loxodonta africana	58.90	58.19	40.66	77.84	41.07	55.75	38.29	29.17	44.15	56.24	41.08	35.13
Macaca fascicularis	60.69	58.40	40.65	83.02	41.27	55.56	38.29	29.96	43.93	56.32	41.19	34.29
Microcebus murinus	59.81	57.52	41.14	80.76	40.41	55.16	38.29	27.78	43.30	55.56	41.00	33.33
Miniopterus natalensis	55.05	57.52	40.95	66.67	47.82	56.75	38.29	48.41	45.50	58.02	40.88	37.58
Monodelphis domestica	46.98	55.43	41.14	44.38	41.01	54.96	38.29	29.76	43.35	55.68	41.23	33.14
Mus musculus	57.17	57.94	41.49	72.08	43.59	56.35	38.10	36.31	46.64	56.05	41.08	42.80
Ochotona princeps	61.27	58.86	41.52	83.43	41.93	55.75	38.29	31.75	46.11	57.09	41.19	40.04
Odocoileus virginianus	59.24	58.48	41.52	77.71	41.73	55.75	38.29	31.15	46.25	57.44	41.60	39.70
Orcinus orca	61.14	57.91	41.33	84.19	41.27	55.95	38.29	29.56	44.19	57.09	41.19	34.29
Ornithorhynchus anatinus	49.27	55.62	41.14	51.05	48.63	56.11	37.88	51.90	44.89	52.60	39.88	42.20
Orycteropus afer	60.00	59.62	41.33	79.05	40.75	54.91	38.08	29.26	45.34	55.94	41.38	38.70
Oryctolagus cuniculus	-	-	-	-	41.53	55.16	38.10	31.35	45.79	56.32	41.00	40.04
Ovis aries	59.24	58.48	41.52	77.71	41.55	55.87	38.37	30.42	45.04	57.06	41.41	36.64
Pan troglodytes	61.27	58.67	40.76	84.38	41.20	55.56	38.29	29.76	43.93	56.13	41.19	34.48
Panthera pardus	61.08	58.29	40.95	84.00	41.20	55.75	38.29	29.56	44.34	56.05	40.88	36.08
Pantholops hodgsonii	59.30	58.48	41.52	77.91	41.47	56.35	38.49	29.56	45.72	58.19	41.37	37.61
Papio anubis	60.62	58.40	40.65	82.82	41.07	55.36	38.29	29.56	44.06	56.32	41.19	34.67
Phascolarctos cinereus	47.87	55.05	41.14	47.43	40.87	55.36	38.29	28.97	43.96	55.56	42.14	34.17
Physeter catodon	61.27	57.71	40.95	85.14	41.49	56.28	38.29	29.92	43.87	57.09	41.00	33.53
Pteropus vampyrus	63.11	58.86	41.33	89.14	42.20	55.95	38.29	32.34	43.95	56.24	41.08	34.55
Rousettus aegyptiacus	63.18	58.86	41.33	89.33	42.13	55.95	38.29	32.14	44.15	56.24	41.08	35.13
Sarcophilus harrisii	46.98	54.48	41.14	45.33	41.01	55.36	38.29	29.37	43.48	55.88	40.66	33.91
Sorex araneus	64.34	59.89	41.62	91.53	42.73	56.55	38.29	33.33	45.87	56.81	40.69	40.12
Trichechus manatus	61.33	59.05	41.52	83.43	41.02	55.11	38.08	29.86	44.63	56.10	40.98	36.79
Tupaia belangeri	63.30	58.97	41.60	89.31	40.74	55.16	38.29	28.77	44.40	57.80	40.22	35.17
Zalophus californianus	59.87	58.67	41.14	79.81	41.34	55.75	38.29	29.96	43.95	56.05	41.08	34.74
Non-mammalians	**49.42**	**55.61**	**40.81**	**51.84**	**43.19**	**54.93**	**38.48**	**36.15**	**45.77**	**55.05**	**40.73**	**41.53**
Acanthisitta chloris	54.52	56.30	41.99	65.27	42.94	55.47	38.37	34.99	-	-	-	-
Alligator mississippiensis	47.37	55.43	40.95	45.71	42.21	55.07	38.17	33.40	45.34	55.88	41.23	38.92
Anas platyrhynchos	47.11	54.67	40.76	45.91	43.25	55.75	38.29	35.71	44.86	56.45	40.43	37.70
Anser cygnoides	47.11	54.67	40.76	45.91	44.30	56.20	41.92	34.77	44.70	54.91	40.66	38.54
Apteryx australis	47.37	55.62	40.76	45.71	42.53	55.95	38.29	33.33	44.70	54.79	40.61	38.70
Chaetura pelagica	50.00	56.01	40.31	53.68	42.69	55.51	38.28	34.27	45.20	54.67	41.44	39.49
Chrysemys picta	48.22	54.77	41.03	48.86	42.20	54.96	38.29	33.33	45.15	55.49	40.85	39.11
Crocodylus porosus	46.79	55.24	40.95	44.19	42.13	54.76	38.29	33.33	45.28	56.26	41.23	38.34
Danio rerio	50.13	55.15	41.60	53.63	51.35	55.05	40.20	58.81	52.24	55.77	40.96	60.00
Gallus gallus	49.18	56.25	40.53	50.76	42.13	55.56	38.29	32.54	43.18	52.30	39.12	38.12
Gavialis gangeticus	47.05	55.43	40.95	44.76	42.02	54.87	38.17	33.00	45.34	56.26	41.23	38.54
Gekko japonicus	53.79	55.26	40.15	65.97	41.62	55.07	38.17	31.61	45.34	54.72	41.04	40.27
Haliaeetus albicilla	46.73	55.05	40.95	44.19	42.73	55.36	38.10	34.72	44.36	53.39	40.62	39.07
Latimeria chalumnae	45.87	54.13	40.88	42.61	42.59	55.36	38.49	33.93	46.69	56.26	41.04	42.78
Lepisosteus oculatus	50.45	55.60	40.35	55.41	45.67	54.65	39.73	42.64	51.47	55.00	40.77	58.65
Nothoprocta perdicaria	48.77	56.17	40.61	49.53	44.09	55.51	38.08	38.68	44.64	53.95	41.04	38.92
Numida meleagris	49.62	55.92	40.65	52.29	42.79	56.35	38.29	33.73	43.31	53.09	39.72	37.13
Oncorhynchus mykiss	57.55	56.79	39.62	76.23	54.36	54.56	39.55	68.97	55.74	55.92	43.80	67.49
Phasianus colchicus	49.05	56.30	40.46	50.38	42.15	55.87	38.17	32.41	42.29	52.31	38.08	36.47
Podarcis muralis	55.11	55.24	41.14	68.95	42.39	54.37	38.10	34.72	45.05	55.02	40.54	39.58
Pogona vitticeps	60.71	57.22	41.26	83.65	41.53	54.96	37.90	31.75	44.57	55.49	41.43	36.80
Python bivittatus	47.79	55.09	40.88	47.41	40.01	53.11	37.88	29.06	44.77	55.11	41.81	37.38
Rhinatrema bivittatum	48.47	55.73	41.41	48.28	41.27	54.17	38.49	31.15	44.99	55.27	40.82	38.87
Rhincodon typus	46.63	54.20	41.41	44.28	43.19	52.34	38.52	38.72	45.65	56.78	37.72	42.44
Struthio camelus	46.92	55.24	40.76	44.76	42.86	55.75	38.29	34.52	45.73	56.26	41.81	39.11
Xenopus laevis	45.89	57.28	40.19	40.19	40.76	53.41	37.35	31.53	44.68	55.13	40.23	38.69
Xenopus tropicalis	46.10	56.67	40.43	41.20	40.30	53.21	37.35	30.32	44.68	54.74	40.62	38.69

Protostoma *PTBP*				
Species	**%C**	**%C1**	**%C2**	**%C3**
Apis mellifera	51.36	54.07	38.70	61.30
Crassostrea gigas	52.66	57.14	41.03	59.80
Drosophila melanogaster	53.84	54.11	41.12	66.28

TABLE C.2 – **Exomic (third position), intronic and flanking regions GC content of *PTBP1*, *PTBP2* and *PTBP3* for the fifteen species used in the genomic context analysis**. For each species, the assembly number to extract the corresponding values is given in parenthesis.

	PTBP1 GC content (%)			*PTBP2* GC content (%)			*PTBP3* GC content (%)		
Species (Assembly number)	GC3	Introns	Flanks	GC3	Introns	Flanks	GC3	Introns	Flanks
Bos taurus (6369068)	77.1	60.1	52.1	31.	34.6	38.3	37.7	36.1	40.7
Canis familiaris (313658)	81.8	62.5	68.4	31.2	32.7	35.8	36.1	35.8	41.5
Danio rerio (482478)	53.0	35.2	34.8	57.7	36.7	36.7	60.2	33.9	33.6
Dasypus novemcinctus (326198)	88.6	67.6	51.3	29.5	33.6	36.3	35.3	36.3	37.2
Equus caballus (286568)	85.0	66.9	66.5	32.1	34.2	37.1	46.6	37.12	42.2
Gallus gallus (6347868)	50.1	42.6	53.1	33.3	36.4	39.9	37.3	36.4	48.3
Gekko japonicus (2693898)	40.4	42.3	43.8	-	-	-	64.6	47.0	49.7
Homo sapiens (8687898)	82.4	60.1	57.3	30.8	35.7	37.1	36.1	37.2	42.2
Latimeria chalumnae (303548)	42.2	34.0	32.9	31.6	32.8	35.0	42.639	36.5	38.4
Loxodonta africana (3288)	-	-	-	29.6	35.2	38.0	35.3	36.3	38.6
Mus musculus (1700338)	72.5	55.7	50.9	37.4	35.7	40.7	42.2	38.0	43.6
Pan troglodytes (5907448)	82.4	60.1	57.0	31.1	35.0	36.9	36.1	38.3	42.7
Pteropus vampyrus (1410408)	89.9	65.6	52.0	32.5	32.2	34.9	34.4	32.4	36.9
Rhincodon typus (4170658)	-	-	-	37.7	39.7	41.0	-	-	-
Xenopus tropicalis (3341798)	40.6	35.8	39.1	31.6	35.2	39.1	39.0	37.5	37.7

TABLE C.3 – $COUSIN_{59}$ **and** CAI **scores of genomic context analysis species.** The reference used is the global CUB of the organism, estimated using all CDSs above 100 amino acids in length.Here, since we calculate the CUB of *PTBP*s against a reference dataset without performing any genomic context analysis, we replaced missing data by individuals *PTBP*s

	PTBP1		***PTBP2***		***PTBP3***	
Species	$COUSIN_{59}$	CAI	$COUSIN_{59}$	CAI	$COUSIN_{59}$	CAI
Bos taurus	1.785	0.787	-0.534	0.618	-0.323	0.642
Canis familiaris	1.976	0.821	-0.499	0.664	-0.207	0.698
Danio rerio	0.655	0.782	1.136	0.774	1.148	0.789
Dasypus novemncinctus	1.958	0.818	-0.57	0.659	-0.277	0.699
Equus caballus	1.831	0.813	-0.489	0.652	0.107	0.698
Gallus gallus	0.764	0.769	-0.006	0.731	0.327	0.745
Gekko japonicus	1.582	0.835	-0.627	0.751	0.199	0.785
Homo sapiens	1.747	0.815	-0.477	0.678	-0.235	0.709
Latimeria chalumnae	0.728	0.818	1.218	0.824	0.4	0.808
Loxodonta africana	2.041	0.821	-0.442	0.704	-0.159	0.74
Mus musculus	1.842	0.829	0.017	0.708	0.154	0.73
Pan troglodytes	1.82	0.81	-0.476	0.682	-0.219	0.712
Pteropus vampyrus	2.112	0.84	-0.427	0.683	-0.287	0.703
Rhincodon typus	0.882	0.791	1.598	0.82	1.163	0.804
Xenopus tropicalis	1.347	0.817	1.559	0.806	1.264	0.808

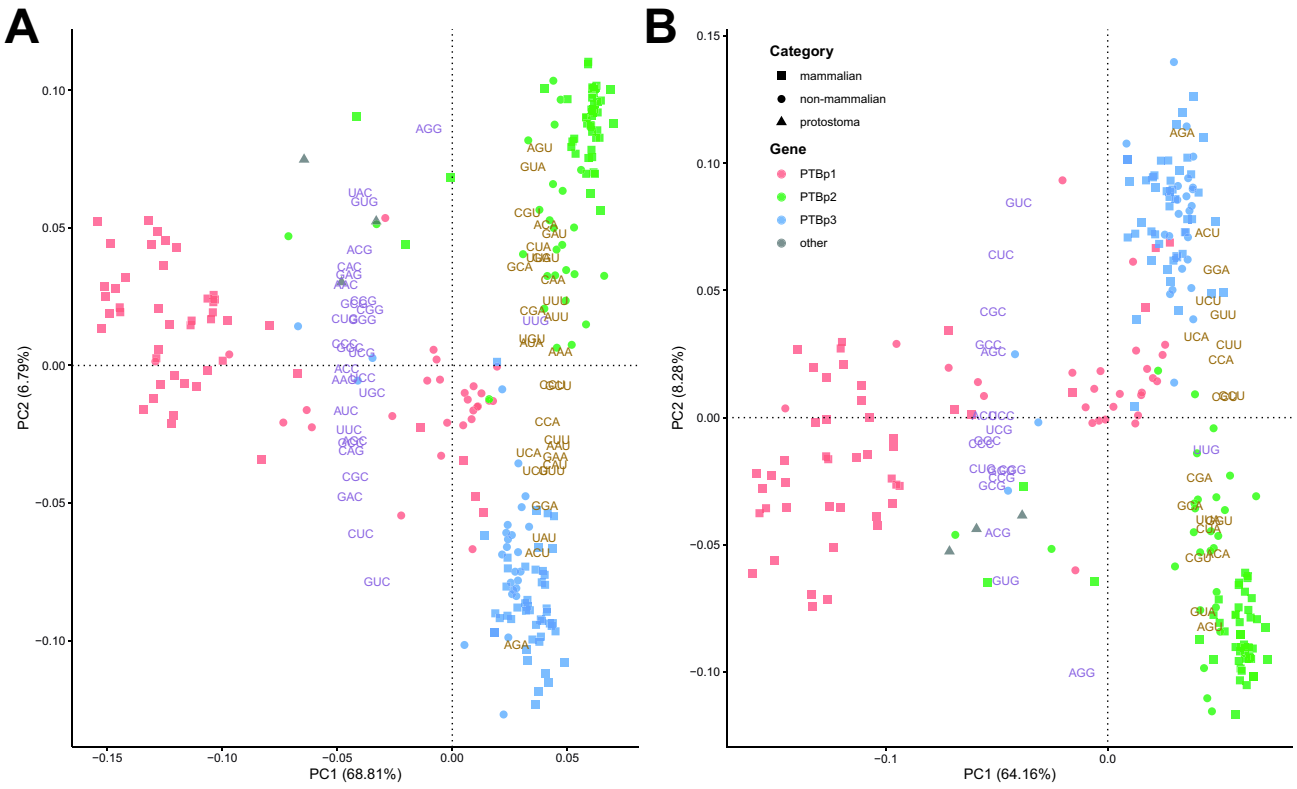

Figure C.2 – **Plot of the two first dimensions of a PCA analysis based on the codon usage preferences of A) all codons B) amino-acids encoded by four codons in *PTBP1*s (red), *PTBP2*s (green), *PTBP3*s (blue) and protostoma (grey) individuals.** Taxonomic information is included as mammals (squares), non-mammals (circles) and protostomates (triangles). The PCA was created using as variables the vectors of 59 positions (representing the relative frequencies of the 59 synonymous codons) for each individual gene. The eigenvalues of the individual codon variables are given by their position on the graph. Each codon variable is identified by its name and by a colour code, purple for GC-ending codons and orange for AT-ending codons. Note the position of the UUG-Leu codon, the sole GC-ending codon clustering with all other AT-ending codons in both panels. The percentage of the total variance explained by each axis is shown in parenthesis.

TABLE C.4 – **Comparison between species tree and subtrees of the nucleotide based maximum likelihood tree.** Each subtree corresponds to a paralog. The K-score compares topological and pairwise distances between trees, while the Robinson-Foulds score compares only topological distances between trees. As the K-score distance is not metric, the analysis has been run both ways.

Reference tree	Comparison tree	K-score	Scale factor	Robinson-Foulds
Nucleotide tree VS species tree				
PTBP1	Species-1	0.759	0.001	42
Species-1	PTBP1	604.403	675.598	42
PTBP2	Species-2	0.762	0.001	24
Species-2	PTBP2	714.262	731.935	24
PTBP3	Species-3	1.700	0.001	28
Species-3	PTBP3	883.463	328.799	28
Nucleotide tree VS Amino acid tree				
PTBP1-AA	PTBP1-NT	0.149	0.197	78
PTBP1-NT	PTBP1-AA	0.666	3.972	78
PTBP2-AA	PTBP2-NT	0.129	0.199	110
PTBP2-NT	PTBP2-AA	0.574	3.917	110
PTBP3-AA	PTBP3-NT	0.380	0.372	40
PTBP3-NT	PTBP3-AA	0.926	2.21	40

TABLE C.5 – **Results of a series of Mantel tests assessing correlation between pairwise nucleotide-based and amino acid-based distance matrices (AA-tree VS NT-tree) and between pairwise nucleotide-based and CUB-based distance matrices (CUB-tree VS NT-tree).** First part is comparison by gene of nucleotide and amino acid based maximum likelihood trees, second part is comparing the same nucleotide based trees by gene against pairwise distances of CUB. Please note the different scales used for the different genes.

AA-tree VS NT-tree	**Observation**	**simulated *p* value**
Mammal		
PTBP1	0.712	0.001
PTBP2	0.558	0.001
PTBP3	0.987	0.001
Non-mammal		
PTBP1	0.929	0.001
PTBP2	0.982	0.001
PTBP3	0.748	0.001
CUB VS NT-tree	**Observation**	**simulated *p* value**
Mammal		
PTBP1	0.908	0.001
PTBP2	0.046	0.274
PTBP3	0.737	0.001
Non-mammal		
PTBP1	0.091	0.273
PTBP3	-0.121	0.73
PTBP3	0.145	0.165

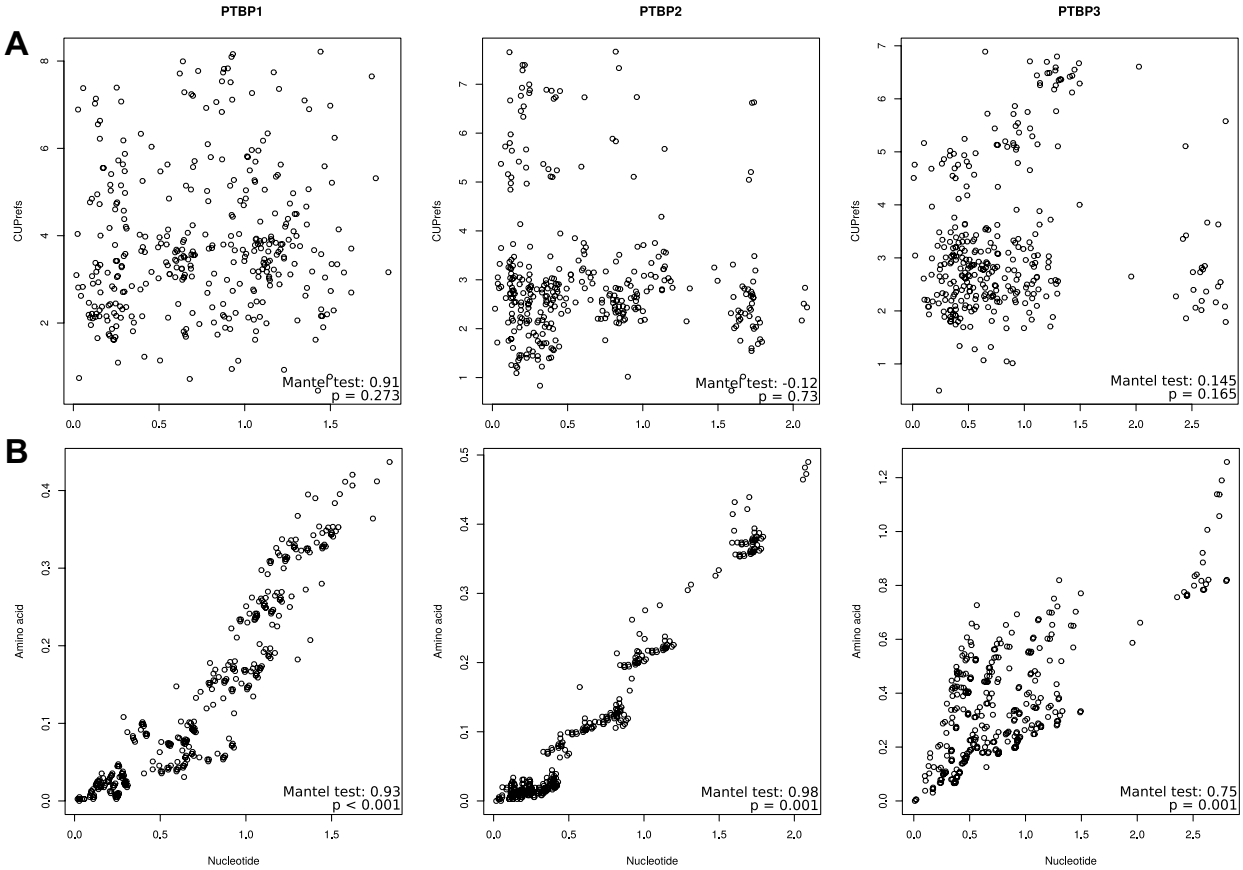

Figure C.3 – **Nucleotide-based pairwise distances against A) CUB and B) amino-acid based pairwise distances for the different non-mammalian *PTBP* orthologs.** The results for a Mantel test assessing the correlation between the corresponding matrices are shown in the inset.

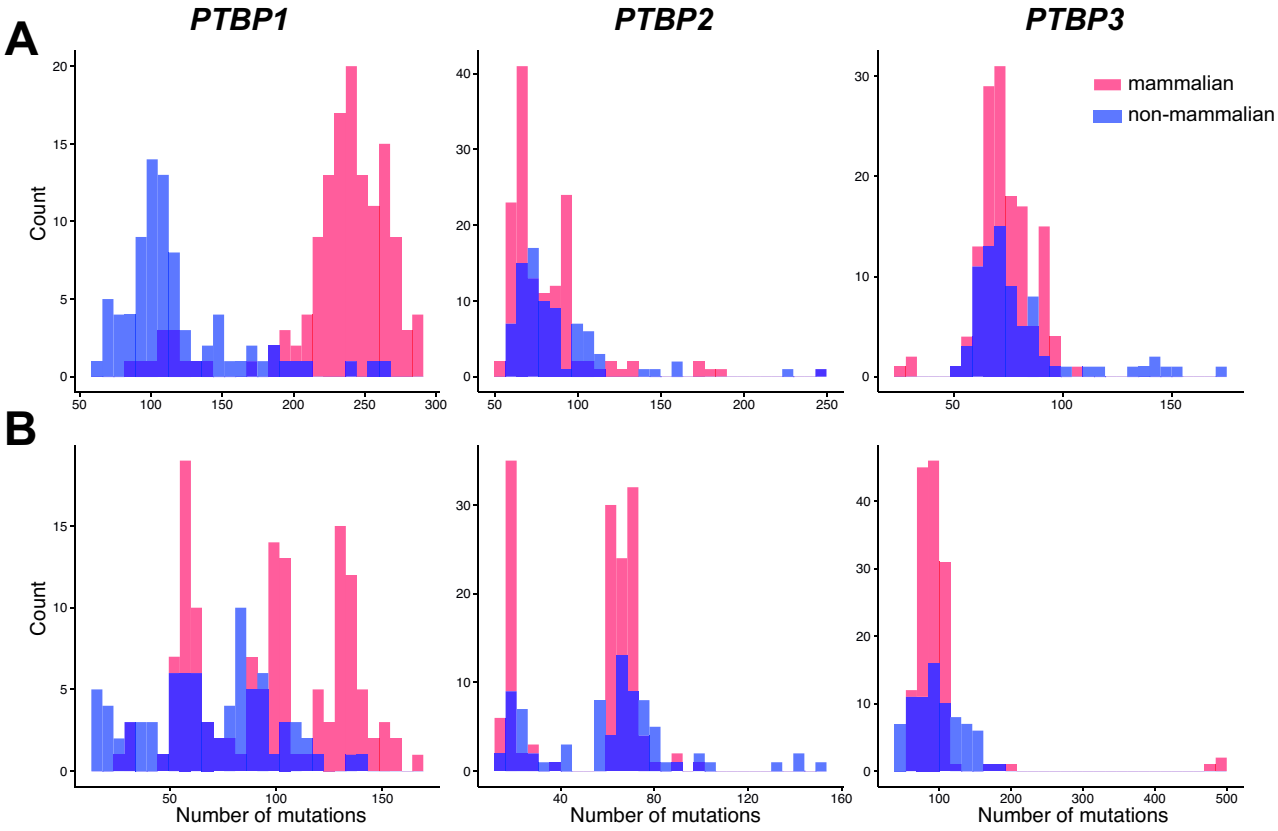

Figure C.4 – **Number of synonymous (A) and non-synonymous mutations (B) for *PTBP1*, *PTBP2* and *PTBP3* mammalian (red) and non-mammalian individuals (blue)**. The informations presented here have been gathered through all individuals studied, with the exception of protostoma species. X-axis represents the number of mutations between a species and its ancestral state and Y-axis represents the frequency of that occurrence.

TABLE C.6 – **Synonymous and non-synonymous matrices for mammalian and non-mammalian *PTBP1*, *PTBP2* and *PTBP3***. The matrices display cumulative counts of substitutions between each individual and their ancestral state. In each matrice, each column represents the count of substitutions from the ancestral stat to individuals.

Synonymous mutations

Mammalians

	PTBP1					***PTBP2***					***PTBP3***			
	A	T	G	C		A	T	G	C		A	T	G	C
A	-	367	566	678	A	-	716	2232	686	A	-	736	1972	1186
T	334	-	109	632	T	774	-	378	1678	T	352	-	175	2380
G	3242	1019	-	631	G	1274	279	-	180	G	996	188	-	223
C	2004	3243	497	-	C	814	1465	67	-	C	538	1077	138	-

Non-mammalians

	PTBP1					***PTBP2***					***PTBP3***			
	A	T	G	C		A	T	G	C		A	T	G	C
A	-	374	1188	626	A	-	401	1036	284	A	-	335	972	502
T	233	-	134	1140	T	446	-	155	836	T	279	-	78	1047
G	924	179	-	208	G	842	253	-	53	G	578	127	-	122
C	452	821	192	-	C	414	906	69	-	C	240	764	63	-

Non synonymous mutations

Mammalians

	PTBP1					***PTBP2***					***PTBP3***			
	A	T	G	C		A	T	G	C		A	T	G	C
A	-	108	814	336	A	-	96	478	94	A	-	327	1286	666
T	241	-	97	37	T	142	-	96	73	T	485	-	215	617
G	1098	270	-	333	G	406	51	-	237	G	1498	147	-	268
C	412	276	364	-	C	120	127	51	-	C	1170	411	196	-

Non-mammalians

	PTBP1					***PTBP2***					***PTBP3***			
	A	T	G	C		A	T	G	C		A	T	G	C
A	-	78	500	274	A	-	75	402	170	A	-	115	640	366
T	119	-	80	76	T	127	-	109	49	T	235	-	130	386
G	302	67	-	140	G	346	48	-	158	G	818	86	-	212
C	228	83	182	-	C	142	116	50	-	C	520	188	132	-

BIBLIOGRAPHIE

[1] G. A. Funk, R. Gosert, P. Comoli, F. Ginevri, et H. H. Hirsch. Polyomavirus BK Replication Dynamics In Vivo and In Silico to Predict Cytopathology and Viral Clearance in Kidney Transplants. *American Journal of Transplantation*, 8(11) :2368–2377, 2008. ISSN 1600-6143. doi : 10.1111/j.1600-6143.2008.02402.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-6143.2008.02402.x. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-6143.2008.02402.x.

[2] Tessa E. F. Quax, Nico J. Claassens, Dieter Söll, et John vanderOost. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, 59(2) :149–161, July 2015. ISSN 1097-2765. doi : 10.1016/j.molcel.2015.05.035. URL http://www.cell.com/molecular-cell/abstract/S1097-2765(15)00402-5.

[3] Burgess Rr, Travers Aa, Dunn Jj, et Bautz Ek. Factor stimulating transcription by RNA polymerase. *Nature*, 221(5175) :43–46, January 1969. ISSN 0028-0836, 1476-4687. doi : 10.1038/221043a0. URL https://europepmc.org/article/med/4882047.

[4] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, et Matthew T. Weirauch. The Human Transcription Factors. *Cell*, 172(4) :650–665, February 2018. ISSN 0092-8674. doi : 10.1016/j.cell.2018.01.029. URL http://www.sciencedirect.com/science/article/pii/S0092867418301065.

[5] Curt M Horvath. STAT proteins and transcriptional responses to extracellular signals. *Trends in Biochemical Sciences*, 25(10) :496–502, October 2000. ISSN 0968-0004. doi : 10.1016/S0968-0004(00)01624-8. URL http://www.sciencedirect.com/science/article/pii/S0968000400016248.

[6] Robert S. Washburn et Max E. Gottesman. Regulation of Transcription Elongation and Termination. *Biomolecules*, 5(2) :1063–1078, May 2015. ISSN 2218-273X. doi : 10.3390/biom5021063. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4496710/.

[7] Harvey Lodish, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore, et James Darnell. Transcription Termination. *Molecular Cell Biology. 4th edition*, 2000. URL https://www.ncbi.nlm.nih.gov/books/NBK21601/. Publisher : W. H. Freeman.

[8] Dianne S. Schwarz et Michael D. Blower. The endoplasmic reticulum : structure, function and response to cellular signaling. *Cellular and Molecular Life Sciences*, 73 :79–94, 2016. ISSN 1420-682X. doi : 10.1007/s00018-015-2052-6. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4700099/.

[9] Avi-ad Avraam Buskila, Shanmugapriya Kannaiah, et Orna Amster-Choder. RNA localization in bacteria. *RNA Biology*, 11(8) :1051–1060, October 2014. ISSN 1547-6286. doi : 10.4161/rna.36135. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4615583/.

[10] K S Kabnick et D E Housman. Determinants that contribute to cytoplasmic stability of human c-fos and beta-globin mRNAs are located at several sites in each mRNA. *Molecular and Cellular Biology*, 8(8) :3244–3250, August 1988. ISSN 0270-7306. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC363556/.

[11] Emmanouil Skoufos et Michle M. Sanders. Regulation of expression of the chicken ovalbumin gene : Interactions between steroid hormones and second messenger systems. *Molecular Endocrinology*, 6(9) :1412–1417, September 1992. ISSN 0888-8809. doi : 10.1210/mend.6.9.1279383. URL https://experts.umn.edu/en/publications/regulation-of-expression-of-the-chicken-ovalbumin-gene-interactio. Publisher : The Endocrine Society.

[12] Edward Yang, Erik van Nimwegen, Mihaela Zavolan, Nikolaus Rajewsky, Mark Schroeder, Marcelo Magnasco, et James E. Darnell. Decay Rates of Human mRNAs : Correlation With Functional Characteristics and Sequence Attributes. *Genome Research*, 13(8) :1863–1872, August 2003. ISSN 1088-9051. doi : 10.1101/gr.1272403. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403777/.

[13] David M. Mauger, B. Joseph Cabral, Vladimir Presnyak, Stephen V. Su, David W. Reid, Brooke Goodman, Kristian Link, Nikhil Khatwani, John Reynders, Melissa J. Moore, et Iain J. McFadyen. mRNA structure regulates protein expression through changes in functional half-life. *Proceedings of the National Academy of Sciences*, 116(48) :24075–24083, November 2019. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1908052116. URL https://www.pnas.org/content/116/48/24075. Publisher : National Academy of Sciences Section : PNAS Plus.

[14] Luba Katz et Christopher B. Burge. Widespread Selection for Local RNA Secondary Structure in Coding Regions of Bacterial Genes. *Genome Research*, 13(9) :2042–2051, September 2003. ISSN 1088-9051. doi : 10.1101/gr.1257503. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403678/.

[15] Devi Mukherjee, Min Gao, J.Patrick O'Connor, Reinout Raijmakers, Ger Pruijn, Carol S. Lutz, et Jeffrey Wilusz. The mammalian exosome mediates the efficient degradation of mRNAs that contain AU-rich elements. *The EMBO Journal*, 21(1-2) :165–174, January 2002. ISSN 0261-4189. doi : 10.1093/emboj/21.1.165. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC125812/.

[16] M F Iademarco, J L Barks, et D C Dean. Regulation of vascular cell adhesion molecule-1 expression by IL-4 and TNF-alpha in cultured endothelial cells. *Journal of Clinical Investigation*, 95(1) :264–271, January 1995. ISSN 0021-9738. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC295423/.

[17] V. Ramakrishnan. Ribosome Structure and the Mechanism of Translation. *Cell*, 108(4) :557–572, February 2002. ISSN 0092-8674. doi : 10.1016/S0092-8674(02)00619-0. URL http://www.sciencedirect.com/science/article/pii/S0092867402006190.

[18] G. E. Palade. A small particulate component of the cytoplasm. *The Journal of Biophysical and Biochemical Cytology*, 1(1) :59–68, January 1955. ISSN 0095-9901. doi : 10.1083/jcb.1.1.59.

[19] Rachel Green et Harry F. Noller. Ribosomes and Translation. *Annual Review of Biochemistry*, 66(1) :679–716, 1997. doi : 10.1146/annurev.biochem.66.1.679. URL https://doi.org/10.1146/annurev.biochem.66.1.679. _eprint : https ://doi.org/10.1146/annurev.biochem.66.1.679.

[20] I G Wool. The Structure and Function of Eukaryotic Ribosomes. *Annual Review of Biochemistry*, 48(1) :719–754, 1979. doi : 10.1146/annurev.bi.48.070179.003443. URL https://doi.org/10.1146/annurev.bi.48.070179.003443. _eprint : https ://doi.org/10.1146/annurev.bi.48.070179.003443.

[21] T. V. Pestova et C. U. T. Hellen. Translation Initiation in Eukaryotes : Factors and Mechanisms. In William J. Lennarz et M. Daniel Lane, editors, *Encyclopedia of Biological Chemistry (Second Edition)*, pages 432–435. Academic Press, Waltham, January 2013. ISBN 978-0-12-378631-9. doi : 10.1016/B978-0-12-378630-2.00482-5. URL http://www.sciencedirect.com/science/article/pii/B9780123786302004825.

[22] Marshall W. Nirenberg et J. Heinrich Matthaei. THE DEPENDENCE OF CELL- FREE PROTEIN SYNTHESIS IN E. COLI UPON NATURALLY OCCURRING OR SYNTHETIC POLYRIBONUCLEOTIDES. *Proceedings of the National Academy of Sciences of the United States of America*, 47(10) :1588–1602, October 1961. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC223178/.

[23] H. G. Khorana, H. Büchi, H. Ghosh, N. Gupta, T. M. Jacob, H. Kössel, R. Morgan, S. A. Narang, E. Ohtsuka, et R. D. Wells. Polynucleotide synthesis and the genetic code. *Cold Spring Harbor Symposia on Quantitative Biology*, 31 :39–49, 1966. ISSN 0091-7451.

[24] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, et Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347) :337–342, May 2011. ISSN 1476-4687. doi : 10.1038/nature10098.

[25] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, et Matthias Selbach. Corrigendum : Global quantification of mammalian gene expression control. *Nature*, 495(7439) :126–127, March 2013. ISSN 1476-4687. doi : 10.1038/nature11848.

[26] Helen Saibil. Chaperone machines for protein folding, unfolding and disaggregation. *Nature reviews. Molecular cell biology*, 14(10) :630–642, October 2013. ISSN 1471-0072. doi : 10.1038/nrm3658. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340576/.

[27] F. H. C. Crick. Codon—anticodon pairing : The wobble hypothesis. *Journal of Molecular Biology*, 19(2) :548–555, August 1966. ISSN 0022-2836. doi : 10.1016/S0022-2836(66)80022-0. URL http://www.sciencedirect.com/science/article/pii/S0022283666800220.

[28] Paul F. Agris. Decoding the genome : a modified view. *Nucleic Acids Research*, 32(1) :223–238, 2004. ISSN 0305-1048. doi : 10.1093/nar/gkh185. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC384350/.

[29] N. Gavini et L. Pulakat. The tRNA species for redundant genetic codons NNU and NNC. A thought on the absence of phenylalanine tRNA with AAA anticodon in Escherichia coli. *The Journal of Biological Chemistry*, 267(4) :2240–2243, February 1992. ISSN 0021-9258.

[30] M. Staehelin. Isoacceptor tRNA's. In E. K. F. Bautz, P. Karlson, et H. Kersten, editors, *Regulation of Transcription and Translation in Eukaryotes*, Colloquium der Gesellschaft für Biologische Chemie 26.–28. April 1973 in Mosbach/Baden, pages 313–321, Berlin, Heidelberg, 1973. Springer. ISBN 978-3-642-65725-2. doi : 10.1007/978-3-642-65725-2_16.

[31] T. Ikemura. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, 2(1) :13–34, January 1985. ISSN 0737-4038. doi : 10.1093/oxfordjournals.molbev.a040335.

[32] R Grantham, C Gautier, M Gouy, R Mercier, et A Pavé. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8(1) :r49–r62, January 1980. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC327256/.

[33] A. Carbone, A. Zinovyev, et F. Képès. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics (Oxford, England)*, 19(16) :2005–2015, November 2003. ISSN 1367-4803.

[34] Vladimir Presnyak, Najwa Alhusaini, Ying-Hsin Chen, Sophie Martin, Nathan Morris, Nicholas Kline, Sara Olson, David Weinberg, Kristian E. Baker, Brenton R. Graveley, et Jeff Coller. Codon optimality is a major determinant of mRNA stability. *Cell*, 160(6) :1111–1124, March 2015. ISSN 0092-8674. doi : 10.1016/j.cell.2015.02.029. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4359748/.

[35] Gareth A. Palidwor, Theodore J. Perkins, et Xuhua Xia. A General Model of Codon Bias Due to GC Mutational Bias. *PLoS ONE*, 5(10), October 2010. ISSN 1932-6203. doi : 10.1371/journal.pone.0013431. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2965080/.

[36] Nicolas Galtier, Camille Roux, Marjolaine Rousselle, Jonathan Romiguier, Emeric Figuet, Sylvain Glémin, Nicolas Bierne, et Laurent Duret. Codon Usage Bias in Animals : Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, 35(5) :1092–1103, May 2018. ISSN 1537-1719. doi : 10.1093/molbev/msy015.

[37] T. Ikemura. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes : a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of Molecular Biology*, 151(3) :389–409, September 1981. ISSN 0022-2836.

[38] J. L. Bennetzen et B. D. Hall. Codon selection in yeast. *Journal of Biological Chemistry*, 257(6) : 3026–3031, March 1982. ISSN 0021-9258, 1083-351X. URL http://www.jbc.org/content/257/6/3026.

[39] Carrie A. Whittle et Cassandra G. Extavour. Expression-Linked Patterns of Codon Usage, Amino Acid Frequency, and Protein Length in the Basally Branching Arthropod Parasteatoda tepidariorum. *Genome Biology and Evolution*, 8(9) :2722–2736, September 2016. doi : 10.1093/gbe/evw068.

URL https://academic.oup.com/gbe/article/8/9/2722/2236211. Publisher : Oxford Academic.

[40] Patricia P. Chan et Todd M. Lowe. GtRNAdb 2.0 : an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Research*, 44(D1) :D184–D189, January 2016. ISSN 0305-1048. doi : 10.1093/nar/gkv1309. URL https://academic.oup.com/nar/article/44/D1/D184/2503100. Publisher : Oxford Academic.

[41] W. Fiers, R. Contreras, F. Duerinck, G. Haegmean, J. Merregaert, W. Min Jou, A. Raeymakers, G. Volckaert, M. Ysebaert, J. Van de Kerckhove, F. Nolf, et M. Van Montagu. A-Protein gene of bacteriophage MS2. *Nature*, 256(5515) :273–278, July 1975. ISSN 1476-4687. doi : 10.1038/256273a0. URL https://www.nature.com/articles/256273a0. Number : 5515 Publisher : Nature Publishing Group.

[42] Argiris Efstratiadis, Fotis C. Kafatos, et Tom Maniatis. The primary structure of rabbit β-globin mRNA as determined from cloned DNA. *Cell*, 10(4) :571–586, April 1977. ISSN 0092-8674. doi : 10.1016/0092-8674(77)90090-3. URL http://www.sciencedirect.com/science/article/pii/0092867477900903.

[43] Fanny Pouyet, Dominique Mouchiroud, Laurent Duret, et Marie Sémon. Recombination, meiotic expression and human codon usage. *eLife*, 6, 2017. ISSN 2050-084X. doi : 10.7554/eLife.27344.

[44] Laurent Duret et Dominique Mouchiroud. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proceedings of the National Academy of Sciences*, 96(8) :4482–4487, April 1999. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.96.8.4482. URL https://www.pnas.org/content/96/8/4482. Publisher : National Academy of Sciences Section : Biological Sciences.

[45] L. E. Orgel. THE MAINTENANCE OF THE ACCURACY OF PROTEIN SYNTHESIS AND ITS RELEVANCE TO AGEING. *Proceedings of the National Academy of Sciences of the United States of America*, 49(4) :517–521, April 1963. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC299893/.

[46] L. E. Post, G. D. Strycharz, M. Nomura, H. Lewis, et P. P. Dennis. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. *Proceedings of the National Academy of Sciences*, 76(4) :1697–1701, April 1979. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.76.4.1697. URL https://www.pnas.org/content/76/4/1697. Publisher : National Academy of Sciences Section : Research Article.

[47] P. M. Sharp et W. H. Li. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3) :1281–1295, February 1987. ISSN 0305-1048.

[48] T. L. Calderone, R. D. Stevens, et T. G. Oas. High-level misincorporation of lysine for arginine at AGA codons in a fusion protein expressed in Escherichia coli. *Journal of Molecular Biology*, 262 (4) :407–412, October 1996. ISSN 0022-2836. doi : 10.1006/jmbi.1996.0524.

[49] S. Kanaya, Y. Yamada, Y. Kudo, et T. Ikemura. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs : gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1) :143–155, September 1999. ISSN 0378-1119. doi : 10.1016/s0378-1119(99)00225-5.

[50] Robin M. Bannerman. Abnormal Haemoglobins in Africa. J. H. P. Jonxis, Ed. Davis, Philadelphia, 1965. xvi + 477 pp. Illus. $20. *Science*, 150(3699) :1017–1017, November 1965. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.150.3699.1017. URL https://science.sciencemag.org/content/150/3699/1017.1. Publisher : American Association for the Advancement of Science Section : Book Reviews.

[51] Patricia Edelmann et Jonathan Gallant. Mistranslation in E. coli. *Cell*, 10(1) :131–137, January 1977. ISSN 0092-8674, 1097-4172. doi : 10.1016/0092-8674(77)90147-7. URL https://www.cell.com/cell/abstract/0092-8674(77)90147-7. Publisher : Elsevier.

[52] L. E. Post et M. Nomura. DNA sequences from the str operon of Escherichia coli. *The Journal of Biological Chemistry*, 255(10) :4660–4666, May 1980. ISSN 0021-9258.

[53] Shixiang Sun, Jingfa Xiao, Huiyong Zhang, et Zhang Zhang. Pangenome Evidence for Higher Codon Usage Bias and Stronger Translational Selection in Core Genes of Escherichia coli. *Frontiers in Microbiology*, 7, 2016. ISSN 1664-302X. doi : 10.3389/fmicb.2016.01180. URL https://www.frontiersin.org/articles/10.3389/fmicb.2016.01180/full. Publisher : Frontiers.

[54] Eduardo P. C. Rocha. Codon usage bias from tRNA's point of view : redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*, 14(11) :2279–2286, November 2004. ISSN 1088-9051. doi : 10.1101/gr.2896904.

[55] Jeffrey R. Powell et Kirstin Dion. Effects of Codon Usage on Gene Expression : Empirical Studies on Drosophila. *Journal of Molecular Evolution*, 80(3-4) :219–226, 2015. ISSN 0022-2844. doi : 10.1007/s00239-015-9675-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4408374/.

[56] H. Akashi. Synonymous Codon Usage in Drosophila Melanogaster : Natural Selection and Translational Accuracy. *Genetics*, 136(3) :927–935, March 1994. ISSN 0016-6731. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205897/.

[57] Anna Chevallier et Jean-Pierre Garel. Studies on tRNA adaptation, tRNA turnover, precursor tRNA and tRNA gene distribution in Bombyx mori by using two-dimensional polyacrylamide gel electrophoresis. *Biochimie*, 61(2) :245–262, April 1979. ISSN 0300-9084. doi : 10.1016/S0300-9084(79)80070-X. URL http://www.sciencedirect.com/science/article/pii/S030090847980070X.

[58] Carrie A. Whittle, Arpita Kulkarni, et Cassandra G. Extavour. Evidence of multifaceted functions of codon usage in translation within the model beetle Tribolium castaneum. *DNA Research*, 26(6) : 473–484, December 2019. doi : 10.1093/dnares/dsz025. URL https://academic.oup.com/dnaresearch/article/26/6/473/5699906. Publisher : Oxford Academic.

[59] K. E. Hastings et C. P. Emerson. Codon usage in muscle genes and liver genes. *Journal of Molecular Evolution*, 19(3-4) :214–218, 1983. ISSN 0022-2844. doi : 10.1007/BF02099968.

[60] Shaokui Yi, Yanhe Li, et Weimin Wang. Selection shapes the patterns of codon usage in three closely related species of genus Misgurnus. *Genomics*, 110(2) :134–142, March 2018. ISSN 0888-7543. doi : 10.1016/j.ygeno.2017.09.004. URL http://www.sciencedirect.com/science/article/pii/S0888754317300812.

[61] Hsiao-Yun Huang et Anita K. Hopper. Multiple Layers of Stress-Induced Regulation in tRNA Biology. *Life*, 6(2), March 2016. ISSN 2075-1729. doi : 10.3390/life6020016. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4931453/.

[62] J Ross. mRNA stability in mammalian cells. *Microbiological Reviews*, 59(3) :423–450, September 1995. ISSN 0146-0749. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC239368/.

[63] Thomas Sweet, Carrie Kovalak, et Jeff Coller. The DEAD-Box Protein Dhh1 Promotes Decapping by Slowing Ribosome Movement. *PLoS Biology*, 10(6), June 2012. ISSN 1544-9173. doi : 10.1371/journal.pbio.1001342. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3373615/.

[64] Thijs Nieuwkoop, Nico J. Claassens, et John van der Oost. Improved protein production and codon optimization analyses in Escherichia coli by bicistronic design. *Microbial Biotechnology*, 12(1) : 173–179, November 2018. ISSN 1751-7915. doi : 10.1111/1751-7915.13332. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6302717/.

[65] Evelina Angov. Codon usage : Nature's roadmap to expression and folding of proteins. *Biotechnology Journal*, 6(6) :650–659, June 2011. ISSN 1860-6768. doi : 10.1002/biot.201000332. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3166658/.

[66] Tamir Tuller et Hadas Zur. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Research*, 43(1) :13–28, January 2015. ISSN 0305-1048. doi : 10.1093/nar/gku1313. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4288200/.

[67] Marilyn Kozak. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. *Nature*, 308(5956) :241–246, March 1984. ISSN 1476-4687. doi : 10.1038/308241a0. URL https://www.nature.com/articles/308241a0. Number : 5956 Publisher : Nature Publishing Group.

[68] Anna E. von Bohlen, Johann Böhm, Ramona Pop, Diana S. Johnson, John Tolmie, Ralf Stücker, Deborah Morris-Rosendahl, et Gerd Scherer. A mutation creating an upstream initiation codon in the SOX9 5' UTR causes acampomelic campomelic dysplasia. *Molecular Genetics & Genomic Medicine*, 5(3) :261–268, March 2017. ISSN 2324-9269. doi : 10.1002/mgg3.282. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5441400/.

[69] Monica Marin. Folding at the rhythm of the rare codon beat. *Biotechnology Journal*, 3(8) :1047–1057, August 2008. ISSN 1860-7314. doi : 10.1002/biot.200800089.

[70] Christine Mordstein, Rosina Savisaar, Robert S. Young, Jeanne Bazile, Lana Talmane, Juliet Luft, Michael Liss, Martin S. Taylor, Laurence D. Hurst, et Grzegorz Kudla. Codon Usage and Splicing Jointly Influence mRNA Localization. *Cell Systems*, 10(4) :351–362.e8, April 2020. ISSN 2405-4712. doi : 10.1016/j.cels.2020.03.001. URL http://www.sciencedirect.com/science/article/pii/S2405471220300806.

[71] Guillermo Lamolle, Mónica Marin, et Fernando Alvarez-Valin. Silent mutations in the gene encoding the p53 protein are preferentially located in conserved amino acid positions and splicing enhancers. *Mutation Research*, 600(1-2) :102–112, August 2006. ISSN 0027-5107. doi : 10.1016/j.mrfmmm.2006.03.004.

[72] Tamir Tuller, Asaf Carmi, Kalin Vestsigian, Sivan Navon, Yuval Dorfan, John Zaborske, Tao Pan, Orna Dahan, Itay Furman, et Yitzhak Pilpel. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell*, 141(2) :344–354, April 2010. ISSN 0092-8674. doi : 10.1016/j.cell.2010.03.031. URL http://www.sciencedirect.com/science/article/pii/S0092867410003193.

[73] Sebastian Pechmann, Justin W. Chartron, et Judith Frydman. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nature Structural & Molecular Biology*, 21(12) :1100–1105, December 2014. ISSN 1545-9985. doi : 10.1038/nsmb.2919.

[74] Etsuko N. Moriyama et Jeffrey R. Powell. Gene length and codon usage bias in Drosophila melanogaster , Saccharomyces cerevisiae and Escherichia coli. *Nucleic Acids Research*, 26 (13) :3188–3193, July 1998. ISSN 0305-1048. doi : 10.1093/nar/26.13.3188. URL https://academic.oup.com/nar/article/26/13/3188/2385889. Publisher : Oxford Academic.

[75] Marc Torrent, Guilhem Chalancon, Natalia S. de Groot, Arthur Wuster, et M. Madan Babu. Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science Signaling*, 11(546), 2018. ISSN 1945-0877. doi : 10.1126/scisignal.aat6409. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6130803/.

[76] Milana Frenkel-Morgenstern, Tamar Danon, Thomas Christian, Takao Igarashi, Lydia Cohen, Ya-Ming Hou, et Lars Juhl Jensen. Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. *Molecular Systems Biology*, 8(1) :572, January 2012. ISSN 1744-4292, 1744-4292. doi : 10.1038/msb.2012.3. URL https://onlinelibrary.wiley.com/doi/abs/10.1038/msb.2012.3.

[77] Gina Cannarozzi, Nicol N. Schraudolph, Mahamadou Faty, Peter von Rohr, Markus T. Friberg, Alexander C. Roth, Pedro Gonnet, Gaston Gonnet, et Yves Barral. A Role for Codon Order in Translation Dynamics. *Cell*, 141(2) :355–367, April 2010. ISSN 0092-8674, 1097-4172. doi : 10.1016/j.cell.2010.02.036. URL https://www.cell.com/cell/abstract/S0092-8674(10)00189-3. Publisher : Elsevier.

[78] J. Robert Coleman, Dimitris Papamichail, Steven Skiena, Bruce Futcher, Eckard Wimmer, et Steffen Mueller. Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. *Science (New York, N.Y.)*, 320(5884) :1784–1787, June 2008. ISSN 0036-8075. doi : 10.1126/science.1155761. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2754401/.

[79] J. Ross Buchan, Lorna S. Aucott, et Ian Stansfield. tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Research*, 34(3) :1015–1027, 2006. ISSN 0305-1048. doi : 10.1093/nar/gkj488. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363775/.

[80] E D Roche et R T Sauer. SsrA-mediated peptide tagging caused by rare codons and tRNA scarcity. *The EMBO Journal*, 18(16) :4579–4589, August 1999. ISSN 0261-4189. doi : 10.1093/emboj/18.16.4579. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1171532/.

[81] Karl J. Fryxell et Won-Jong Moon. CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content. *Molecular Biology and Evolution*, 22(3) :650–658, March 2005. ISSN 0737-4038. doi : 10.1093/molbev/msi043. URL https://doi.org/10.1093/molbev/msi043.

[82] Miho M. Suzuki et Adrian Bird. DNA methylation landscapes : provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6) :465–476, June 2008. ISSN 1471-0064. doi : 10.1038/nrg2341. URL https://www.nature.com/articles/nrg2341. Number : 6 Publisher : Nature Publishing Group.

[83] Benoît Aliaga, Ingo Bulla, Gabriel Mouahid, David Duval, et Christoph Grunau. Universality of the DNA methylation codes in Eucaryotes. *Scientific Reports*, 9(1) :173, January 2019. ISSN 2045-2322. doi : 10.1038/s41598-018-37407-8. URL https://www.nature.com/articles/s41598-018-37407-8. Number : 1 Publisher : Nature Publishing Group.

[84] Fiona Tulloch, Nicky J Atkinson, David J Evans, Martin D Ryan, et Peter Simmonds. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. *eLife*, 3, 2014. ISSN 2050-084X. doi : 10.7554/eLife.04531. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383024/.

[85] Valerie Odon, Jelke J. Fros, Niluka Goonawardane, Isabelle Dietrich, Ahmad Ibrahim, Kinda Alshaikhahmed, Dung Nguyen, et Peter Simmonds. The role of ZAP and OAS3/RNAseL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides. *Nucleic Acids Research*, 47(15) :8061–8083, September 2019. ISSN 0305-1048. doi : 10.1093/nar/gkz581. URL https://academic.oup.com/nar/article/47/15/8061/5528741. Publisher : Oxford Academic.

[86] Patricia Cortazzo, Carlos Cerveñansky, Mónica Marı ?n, Claude Reiss, Ricardo Ehrlich, et Atilio Deana. Silent mutations affect in vivo protein folding in Escherichia coli. *Biochemical and Biophysical Research Communications*, 293(1) :537–541, April 2002. ISSN 0006-291X. doi : 10.1016/S0006-291X(02)00226-7. URL http://www.sciencedirect.com/science/article/pii/S0006291X02002267.

[87] Tanya Crombie, John P. Boyle, John R. Coggins, et Alistair J. P. Brown. The Folding of the Bifunctional TRP3 Protein in Yeast is Influenced by a Translational Pause which Lies in a Region of Structural Divergence with Escherichia coli Indoleglycerol-Phosphate Synthase. *European Journal of Biochemistry*, 226(2) :657–664, 1994. ISSN 1432-1033. doi : 10.1111/j.1432-1033.1994.tb20093.x. URL https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1432-1033.1994.tb20093.x. _eprint : https ://febs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1432-1033.1994.tb20093.x.

[88] Sebastian Pechmann et Judith Frydman. Evolutionary conservation of codon optimality reveals hidden signatures of co-translational folding. *Nature structural & molecular biology*, 20(2) :237–243, February 2013. ISSN 1545-9993. doi : 10.1038/nsmb.2466. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3565066/.

[89] Kalpit Shah, Yi Cheng, Brian Hahn, Robert Bridges, Neil Bradbury, et David M. Mueller. Synonymous Codon Usage Affects the Expression of Wild Type and F508del CFTR. *Journal of molecular biology*, 427(6 0 0) :1464–1479, March 2015. ISSN 0022-2836. doi : 10.1016/j.jmb.2015.02.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4355305/.

[90] H. Akashi. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics*, 139(2) :1067–1076, February 1995. ISSN 0016-6731.

[91] Philipp W. Messer. Measuring the Rates of Spontaneous Mutation From Deep and Large-Scale Polymorphism Data. *Genetics*, 182(4) :1219–1232, August 2009. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.109.105692. URL https://www.genetics.org/content/182/4/1219. Publisher : Genetics Section : Investigations.

[92] Ruth Hershberg et Dmitri A. Petrov. Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLoS Genetics*, 6(9), September 2010. ISSN 1553-7390. doi : 10.1371/journal.pgen.1001115. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2936535/.

[93] Penelope R Haddrill et Brian Charlesworth. Non-neutral processes drive the nucleotide composition of non-coding sequences in Drosophila. *Biology Letters*, 4(4) :438–441, August 2008. doi : 10.1098/rsbl.2008.0174. URL https://royalsocietypublishing.org/doi/full/10.1098/rsbl.2008.0174. Publisher : Royal Society.

[94] Jérôme Bourret, Samuel Alizon, et Ignacio G. Bravo. COUSIN (COdon Usage Similarity INdex) : A Normalized Measure of Codon Usage Preferences. *Genome Biology and Evolution*, 11(12) : 3523–3528, December 2019. doi : 10.1093/gbe/evz262. URL https://academic.oup.com/gbe/article/11/12/3523/5652094. Publisher : Oxford Academic.

[95] Hongan Long, Way Sung, Sibel Kucukyildirim, Emily Williams, Samuel F. Miller, Wanfeng Guo, Caitlyn Patterson, Colin Gregory, Chloe Strauss, Casey Stone, Cécile Berne, David Kysela, William R. Shoemaker, Mario E. Muscarella, Haiwei Luo, Jay T. Lennon, Yves V. Brun, et Michael Lynch. Evolutionary determinants of genome-wide nucleotide composition. *Nature Ecology & Evolution*, 2(2) :237–240, February 2018. ISSN 2397-334X. doi : 10.1038/s41559-017-0425-y. URL https://www.nature.com/articles/s41559-017-0425-y. Number : 2 Publisher : Nature Publishing Group.

[96] Nicolas Galtier et Laurent Duret. Adaptation or biased gene conversion ? Extending the null hypothesis of molecular evolution. *Trends in genetics : TIG*, 23(6) :273–277, June 2007. ISSN 0168-9525. doi : 10.1016/j.tig.2007.03.011.

[97] Dmitri A. Petrov et Daniel L. Hartl. Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proceedings of the National Academy of Sciences*, 96(4) :1475–1479, February 1999. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.96.4.1475. URL https://www.pnas.org/content/96/4/1475. Publisher : National Academy of Sciences Section : Biological Sciences.

[98] Peter M.J. Burgers et Thomas A. Kunkel. Eukaryotic DNA Replication Fork. *Annual Review of Biochemistry*, 86(1) :417–438, 2017. doi : 10.1146/annurev-biochem-061516-044709. URL https://doi.org/10.1146/annurev-biochem-061516-044709. _eprint : https ://doi.org/10.1146/annurev-biochem-061516-044709.

[99] Martin A. M. Reijns, Harriet Kemp, James Ding, Sophie Marion de Procé, Andrew P. Jackson, et Martin S. Taylor. Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, 518(7540) :502–506, February 2015. ISSN 1476-4687. doi : 10.1038/nature14183. URL https://www.nature.com/articles/nature14183. Number : 7540 Publisher : Nature Publishing Group.

[100] A. Beletskii et Ashok S. Bhagwat. Transcription-induced mutations : Increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 93(24) :13919–13924, November 1996. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC19468/.

[101] Josep M. Comeron. Selective and Mutational Patterns Associated With Gene Expression in Humans : Influences on Synonymous Composition and Intron Presence. *Genetics*, 167(3) :1293–1304, July 2004. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.104.026351. URL https://www.genetics.org/content/167/3/1293. Publisher : Genetics Section : Investigations.

[102] M. Gardiner-Garden et M. Frommer. CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2) :261–282, July 1987. ISSN 0022-2836. doi : 10.1016/0022-2836(87)90689-9. URL http://www.sciencedirect.com/science/article/pii/0022283687906899.

[103] Gabriel Marais. Biased gene conversion : implications for genome and sex evolution. *Trends in Genetics*, 19(6) :330–338, June 2003. ISSN 0168-9525. doi : 10.1016/S0168-9525(03)00116-1. URL https://www.cell.com/trends/genetics/abstract/S0168-9525(03)00116-1. Publisher : Elsevier.

[104] Ran Li, Emmanuelle Bitoun, Nicolas Altemose, Robert W. Davies, Benjamin Davies, et Simon R. Myers. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nature Communications*, 10(1) :3900, August 2019. ISSN 2041-1723. doi : 10.1038/s41467-019-11675-y. URL https://www.nature.com/articles/s41467-019-11675-y. Number : 1 Publisher : Nature Publishing Group.

[105] Yann Lesecque, Dominique Mouchiroud, et Laurent Duret. GC-Biased Gene Conversion in Yeast Is Specifically Associated with Crossovers : Molecular Mechanisms and Evolutionary Significance. *Molecular Biology and Evolution*, 30(6) :1409–1419, June 2013. ISSN 0737-4038. doi : 10.1093/molbev/mst056. URL https://doi.org/10.1093/molbev/mst056.

[106] Laurent Duret, Adam Eyre-Walker, et Nicolas Galtier. A new perspective on isochore evolution. *Gene*, 385 :71–74, December 2006. ISSN 0378-1119. doi : 10.1016/j.gene.2006.04.030.

[107] N. Galtier, G. Piganeau, D. Mouchiroud, et L. Duret. GC-Content Evolution in Mammalian Genomes : The Biased Gene Conversion Hypothesis. *Genetics*, 159(2) :907–911, October 2001. ISSN 0016-6731, 1943-2631. URL https://www.genetics.org/content/159/2/907. Publisher : Genetics Section : Letter to the Editor.

[108] Mart Krupovic, Valerian V. Dolja, et Eugene V. Koonin. Origin of viruses : primordial replicators recruiting capsids from hosts. *Nature Reviews Microbiology*, 17(7) :449–458, July 2019. ISSN 1740-1534. doi : 10.1038/s41579-019-0205-6. URL https://www.nature.com/articles/s41579-019-0205-6. Number : 7 Publisher : Nature Publishing Group.

[109] D Baltimore. Expression of animal virus genomes. *Bacteriological Reviews*, 35(3) :235–241, September 1971. ISSN 0005-3678. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC378387/.

[110] Harvey Lodish, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore, et James Darnell. Viruses : Structure, Function, and Uses. *Molecular Cell Biology. 4th edition*, 2000. URL https://www.ncbi.nlm.nih.gov/books/NBK21523/. Publisher : W. H. Freeman.

[111] V.Gregory Chinchar. REPLICATION OF VIRUSES. *Encyclopedia of Virology*, pages 1471–1478, 1999. doi : 10.1006/rwvi.1999.0245. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149704/.

[112] Eli Goz, Oriah Mioduser, Alon Diament, et Tamir Tuller. Evidence of translation efficiency adaptation of the coding regions of the bacteriophage lambda. *DNA Research*, 24(4) :333–342, August 2017. ISSN 1340-2838. doi : 10.1093/dnares/dsx005. URL https://academic.oup.com/dnaresearch/article/24/4/333/3058514. Publisher : Oxford Academic.

[113] Sheng-Lin Shi, Yi-Ren Jiang, Yan-Qun Liu, Run-Xi Xia, et Li Qin. Selective pressure dominates the synonymous codon usage in parvoviridae. *Virus Genes*, 46(1) :10–19, February 2013. ISSN 1572-994X. doi : 10.1007/s11262-012-0818-6.

[114] Gareth M. Jenkins et Edward C. Holmes. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Research*, 92(1) :1–7, March 2003. ISSN 0168-1702.

[115] Cara Carthel Burns, Jing Shaw, Ray Campagnoli, Jaume Jorba, Annelet Vincent, Jacqueline Quay, et Olen Kew. Modulation of Poliovirus Replicative Fitness in HeLa Cells by Deoptimization of Synonymous Codon Usage in the Capsid Region. *Journal of Virology*, 80(7) :3259–3272, April 2006. ISSN 0022-538X, 1098-5514. doi : 10.1128/JVI.80.7.3259-3272.2006. URL https://jvi.asm.org/content/80/7/3259. Publisher : American Society for Microbiology Journals Section : GENETIC DIVERSITY AND EVOLUTION.

[116] Stephanie Michely, Eve Toulza, Lucie Subirana, Uwe John, Valérie Cognat, Laurence Maréchal-Drouard, Nigel Grimsley, Hervé Moreau, et Gwenaël Piganeau. Evolution of Codon Usage in the Smallest Photosynthetic Eukaryotes and Their Giant Viruses. *Genome Biology and Evolution*, 5 (5) :848–859, 2013. ISSN 1759-6653. doi : 10.1093/gbe/evt053. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3673656/.

[117] Julius B. Lucks, David R. Nelson, Grzegorz R. Kudla, et Joshua B. Plotkin. Genome Landscapes and Bacteriophage Codon Usage. *PLoS Computational Biology*, 4(2), February 2008. ISSN 1553-734X. doi : 10.1371/journal.pcbi.1000001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2266997/.

[118] Oriah Mioduser, Eli Goz, et Tamir Tuller. Significant differences in terms of codon usage bias between bacteriophage early and late genes : a comparative genomics analysis. *BMC Genomics*, 18(1) :866, November 2017. ISSN 1471-2164. doi : 10.1186/s12864-017-4248-7. URL https://doi.org/10.1186/s12864-017-4248-7.

[119] Kajal Kumar Biswas, Supratik Palchoudhury, Prosenjit Chakraborty, Utpal K. Bhattacharyya, Dilip K. Ghosh, Palash Debnath, Chandrika Ramadugu, Manjunath L. Keremane, Ravi K. Khetarpal, et Richard F. Lee. Codon Usage Bias Analysis of Citrus tristeza virus : Higher Codon Adaptation to Citrus reticulata Host. *Viruses*, 11(4), April 2019. ISSN 1999-4915. doi : 10.3390/v11040331. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6521185/.

[120] Azeem Mehmood Butt, Izza Nasrullah, Raheel Qamar, et Yigang Tong. Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerging Microbes & Infections*, 5(10) :e107, October 2016. ISSN 2222-1751. doi : 10.1038/emi.2016.106. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5117728/.

[121] Azeem Mehmood Butt, Izza Nasrullah, et Yigang Tong. Genome-Wide Analysis of Codon Usage and Influencing Factors in Chikungunya Viruses. *PLoS ONE*, 9(3), March 2014. ISSN 1932-6203. doi : 10.1371/journal.pone.0090905. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3942501/.

[122] Izza Nasrullah, Azeem M. Butt, Shifa Tahir, Muhammad Idrees, et Yigang Tong. Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. *BMC Evolutionary Biology*, 15(1) :174, August 2015. ISSN 1471-2148. doi : 10.1186/s12862-015-0456-4. URL https://doi.org/10.1186/s12862-015-0456-4.

[123] Juan Cristina, Pilar Moreno, Gonzalo Moratorio, et Héctor Musto. Genome-wide analysis of codon usage bias in Ebolavirus. *Virus research*, 196C :87–93, November 2014. doi : 10.1016/j.virusres.2014.11.005.

[124] Emily HM Wong, David K Smith, Raul Rabadan, Malik Peiris, et Leo LM Poon. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evolutionary Biology*, 10 :253, August 2010. ISSN 1471-2148. doi : 10.1186/1471-2148-10-253. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2933640/.

[125] Glòria Sánchez, Albert Bosch, et Rosa M. Pintó. Genome Variability and Capsid Structural Constraints of Hepatitis A Virus. *Journal of Virology*, 77(1) :452–459, January 2003. ISSN 0022-538X. doi : 10.1128/JVI.77.1.452-459.2003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC140588/.

[126] Siddiq Ur Rahman, Xiaoting Yao, Xiangchen Li, Dekun Chen, et Shiheng Tao. Analysis of codon usage bias of Crimean-Congo hemorrhagic fever virus and its adaptation to hosts. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 58 :1–16, 2018. ISSN 1567-7257. doi : 10.1016/j.meegid.2017.11.027.

[127] Ming-ren Ma, Xiao-qin Ha, Hui Ling, Mei-liang Wang, Fang-xin Zhang, Shang-di Zhang, Ge Li, et Wei Yan. The characteristics of the synonymous codon usage in hepatitis B virus and the effects of host on the virus in codon usage pattern. *Virology Journal*, 8 :544, December 2011. ISSN

1743-422X. doi : 10.1186/1743-422X-8-544. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3287100/.

[128] Sabyasachi Das, Sandip Paul, et Chitra Dutta. Synonymous codon usage in adenoviruses : influence of mutation, selection and protein hydropathy. *Virus Research*, 117(2) :227–236, May 2006. ISSN 0168-1702. doi : 10.1016/j.virusres.2005.10.007.

[129] Ignacio G. Bravo et Martin Müller. Codon usage in papillomavirus genes : practical and functional aspects. *Papillomavirus Report*, 16(2) :63–72, March 2005. ISSN 0957-4190. doi : 10.1179/095741905X24996. URL https://doi.org/10.1179/095741905X24996. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1179/095741905X24996.

[130] S Karlin, B E Blaisdell, et G A Schachtel. Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus : data and hypotheses. *Journal of Virology*, 64(9) :4264–4273, September 1990. ISSN 0022-538X. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC247892/.

[131] Feng Chen, Peng Wu, Shuyun Deng, Heng Zhang, Yutong Hou, Zheng Hu, Jianzhi Zhang, Xiaoshu Chen, et Jian-Rong Yang. Dissimilation of synonymous codon usage bias in virus–host coevolution due to translational selection. *Nature Ecology & Evolution*, 4(4) :589–600, April 2020. ISSN 2397-334X. doi : 10.1038/s41559-020-1124-7. URL https://www.nature.com/articles/s41559-020-1124-7. Number : 4 Publisher : Nature Publishing Group.

[132] Nicholas Arpaia et Gregory M. Barton. Toll-like Receptors : Key Players in Antiviral Immunity. *Current Opinion in Virology*, 1(6) :447–454, December 2011. ISSN 1879-6257. doi : 10.1016/j.coviro.2011.10.006. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3311989/.

[133] Yasukazu Nakamura, Takashi Gojobori, et Toshimichi Ikemura. Codon usage tabulated from international DNA sequence databases : status for the year 2000. *Nucleic Acids Research*, 28(1) : 292, January 2000. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102460/.

[134] M. A. Freire-Picos, M. I. González-Siso, E. Rodríguez-Belmonte, A. M. Rodríguez-Torres, E. Ramil, et M. E. Cerdán. Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes. *Gene*, 139(1) :43–49, February 1994. ISSN 0378-1119.

[135] Mario dos Reis, Renos Savva, et Lorenz Wernisch. Solving the riddle of codon usage preferences : a test for translational selection. *Nucleic Acids Research*, 32(17) :5036–5044, 2004. ISSN 0305-1048. doi : 10.1093/nar/gkh834. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC521650/.

[136] Soohyun Lee, Seyeon Weon, Sooncheol Lee, et Changwon Kang. Relative Codon Adaptation Index, a Sensitive Measure of Codon Usage Bias. *Evolutionary Bioinformatics Online*, 6 :47–55, May 2010. ISSN 1176-9343. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2880845/.

[137] D. C. Shields, P. M. Sharp, D. G. Higgins, et F. Wright. "Silent" sites in Drosophila genes are not neutral : evidence of selection among synonymous codons. *Molecular Biology and Evolution*, 5 (6) :704–716, November 1988. ISSN 0737-4038.

[138] B. R. Morton. Chloroplast DNA codon use : evidence for selection at the psb A locus based on tRNA availability. *Journal of Molecular Evolution*, 37(3) :273–280, September 1993. ISSN 0022-2844.

[139] A. O. Urrutia et L. D. Hurst. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, 159(3) :1191–1199, November 2001. ISSN 0016-6731.

[140] Xiu-Feng Wan, Dong Xu, Andris Kleinhofs, et Jizhong Zhou. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evolutionary Biology*, 4 :19, 2004. ISSN 1471-2148. doi : 10.1186/1471-2148-4-19. URL http://dx.doi.org/10.1186/1471-2148-4-19.

[141] Zhang Zhang, Jun Li, Peng Cui, Feng Ding, Ang Li, Jeffrey P. Townsend, et Jun Yu. Codon Deviation Coefficient : a novel measure for estimating codon usage bias and its statistical significance. *BMC Bioinformatics*, 13 :43, 2012. ISSN 1471-2105. doi : 10.1186/1471-2105-13-43. URL http://dx.doi.org/10.1186/1471-2105-13-43.

[142] F. Wright. The 'effective number of codons' used in a gene. *Gene*, 87(1) :23–29, March 1990. ISSN 0378-1119.

[143] John A. Novembre. Accounting for background nucleotide composition when measuring codon usage bias. *Molecular Biology and Evolution*, 19(8) :1390–1394, August 2002. ISSN 0737-4038.

[144] Anders Fuglsang. The 'effective number of codons' revisited. *Biochemical and Biophysical Research Communications*, 317(3) :957–964, May 2004. ISSN 0006-291X. doi : 10.1016/j.bbrc.2004.03.138.

[145] Siddhartha Sankar Satapathy, Ajit Kumar Sahoo, Suvendra Kumar Ray, et Tapash Chandra Ghosh. Codon degeneracy and amino acid abundance influence the measures of codon usage bias : improved Nc (N ?c) and ENCprime (N ?'c) measures. *Genes to Cells*, 22(3) :277–283, March 2017. ISSN 1365-2443. doi : 10.1111/gtc.12474. URL http://onlinelibrary.wiley.com/doi/10.1111/gtc.12474/abstract.

[146] Michael C. Angellotti, Shafquat B. Bhuiyan, Guorong Chen, et Xiu-Feng Wan. CodonO : codon usage bias analysis within and across genomes. *Nucleic Acids Research*, 35(suppl_2) :W132–W136, July 2007. ISSN 0305-1048. doi : 10.1093/nar/gkm392. URL https://academic.oup.com/nar/article/35/suppl_2/W132/2923883/CodonO-codon-usage-bias-analysis-within-and-across.

[147] J. M. Comeron et M. Aguadé. An evaluation of measures of synonymous codon usage bias. *Journal of Molecular Evolution*, 47(3) :268–274, September 1998. ISSN 0022-2844.

[148] Alexander Roth, Maria Anisimova, et Gina M. Cannarozzi. *Measuring codon usage bias*. February 2012. ISBN 978-0-19-960116-5. URL https://www.researchgate.net/publication/230857012_Measuring_codon_usage_bias.

[149] Damien Paulet, Alexandre David, et Eric Rivals. Ribo-seq enlightens codon usage bias. *DNA Research*, 24(3) :303–210, June 2017. ISSN 1340-2838, 1756-1663. doi : 10.1093/dnares/dsw062. URL https://academic.oup.com/dnaresearch/article-lookup/doi/10.1093/dnares/dsw062.

[150] John Peden et Paul Sharp. Correspondence Analysis of Codon Usage, 2005. URL http://codonw.sourceforge.net/.

[151] Andreas Grote, Karsten Hiller, Maurice Scheer, Richard Münch, Bernd Nörtemann, Dietmar C. Hempel, et Dieter Jahn. JCat : a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Research*, 33(Web Server issue) :W526–W531, July 2005. ISSN 0305-1048. doi : 10.1093/nar/gki376. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1160137/.

[152] Pere Puigbò, Ignacio G. Bravo, et Santiago Garcia-Vallve. CAIcal : A combined set of tools to assess codon usage adaptation. *Biology Direct*, 3 :38, 2008. ISSN 1745-6150. doi : 10.1186/1745-6150-3-38. URL http://dx.doi.org/10.1186/1745-6150-3-38.

[153] Umashankar Vetrivel, Vijayakumar Arunkumar, et Sudarsanam Dorairaj. ACUA : A software tool for automated codon usage analysis. *Bioinformation*, 2(2) :62–63, October 2007. ISSN 0973-2063. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2174420/.

[154] P. Rice, I. Longden, et A. Bleasby. EMBOSS : the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, 16(6) :276–277, June 2000. ISSN 0168-9525.

[155] Fran Supek et Kristian Vlahovicek. INCA : synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics (Oxford, England)*, 20(14) :2329–2330, September 2004. ISSN 1367-4803. doi : 10.1093/bioinformatics/bth238.

[156] Sarah M. Richardson, Sarah J. Wheelan, Robert M. Yarrington, et Jef D. Boeke. GeneDesign : Rapid, automated design of multikilobase synthetic genes. *Genome Research*, 16(4) :550–556, April 2006. ISSN 1088-9051. doi : 10.1101/gr.4431306. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1457031/.

[157] Pere Puigbò, Eduard Guzmán, Antoni Romeu, et Santiago Garcia-Vallvé. OPTIMIZER : a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Research*, 35(suppl_2) :W126–W131, July 2007. ISSN 0305-1048. doi : 10.1093/nar/gkm219. URL https://academic.oup.com/nar/article/35/suppl_2/W126/2920747/OPTIMIZER-a-web-server-for-optimizing-the-codon.

[158] Pere Puigbò, Ignacio G. Bravo, et Santiago Garcia-Vallvé. E-CAI : a novel server to estimate an expected value of Codon Adaptation Index (eCAI). *BMC Bioinformatics*, 9 :65, 2008. ISSN 1471-2105. doi : 10.1186/1471-2105-9-65. URL http://dx.doi.org/10.1186/1471-2105-9-65.

[159] Ludwik Gross. A Filterable Agent, Recovered from Ak Leukemic Extracts, Causing Salivary Gland Carcinomas in C3H Mice. *Proceedings of the Society for Experimental Biology and Medicine*, 83(2) :414–421, June 1953. ISSN 0037-9727. doi : 10.3181/00379727-83-20376. URL https://journals.sagepub.com/doi/abs/10.3181/00379727-83-20376. Publisher : SAGE Publications.

[160] Sarah E. Stewart, Bernice E. Eddy, Alice M. Gochenour, Ninette G. Borgese, et George E. Grubbs. The induction of neoplasms with a substance released from mouse tumors by tissue culture. *Virology*, 3(2) :380–400, April 1957. ISSN 0042-6822. doi : 10.1016/0042-6822(57)90100-9. URL http://www.sciencedirect.com/science/article/pii/0042682257901009.

[161] Ugo Moens, Sébastien Calvignac-Spencer, Chris Lauber, Torbjörn Ramqvist, Mariet C. W. Feltkamp, Matthew D. Daugherty, Ernst J. Verschoor, et Bernhard Ehlers. ICTV Virus Taxonomy Profile : Polyomaviridae. *The Journal of General Virology*, 98(6) :1159–1160, June 2017. ISSN 0022-1317. doi : 10.1099/jgv.0.000839. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5656788/.

[162] B. H. Sweet et M. R. Hilleman. The vacuolating virus, S.V. 40. *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.)*, 105 :420–427, November 1960. ISSN 0037-9727. doi : 10.3181/00379727-105-26128.

[163] Lin Liu. Fields Virology, 6th Edition. *Clinical Infectious Diseases*, 59(4) :613–613, August 2014. ISSN 1058-4838. doi : 10.1093/cid/ciu346. URL https://academic.oup.com/cid/article/59/4/613/2895607. Publisher : Oxford Academic.

[164] C. Cicala, F. Pompetti, et M. Carbone. SV40 induces mesotheliomas in hamsters. *The American Journal of Pathology*, 142(5) :1524–1533, May 1993. ISSN 0002-9440. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1886912/.

[165] Regis A. Vilchez et Janet S. Butel. Emergent Human Pathogen Simian Virus 40 and Its Role in Cancer. *Clinical Microbiology Reviews*, 17(3) :495–508, July 2004. ISSN 0893-8512. doi : 10.1128/CMR.17.3.495-508.2004. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC452549/.

[166] Michael Fitzpatrick. The Cutter Incident : How America's First Polio Vaccine Led to a Growing Vaccine Crisis. *Journal of the Royal Society of Medicine*, 99(3) :156, March 2006. ISSN 0141-0768. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1383764/.

[167] B. L. Padgett, D. L. Walker, G. M. ZuRhein, R. J. Eckroade, et B. H. Dessel. Cultivation of papova-like virus from human brain with progressive multifocal leucoencephalopathy. *Lancet (London, England)*, 1(7712) :1257–1260, June 1971. ISSN 0140-6736. doi : 10.1016/s0140-6736(71)91777-6.

[168] S. D. Gardner, A. M. Field, D. V. Coleman, et B. Hulme. New human papovavirus (B.K.) isolated from urine after renal transplantation. *Lancet (London, England)*, 1(7712) :1253–1257, June 1971. ISSN 0140-6736. doi : 10.1016/s0140-6736(71)91776-4.

[169] Huichen Feng, Masahiro Shuda, Yuan Chang, et Patrick S. Moore. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science (New York, N.Y.)*, 319(5866) :1096–1100, February 2008. ISSN 1095-9203. doi : 10.1126/science.1152586.

[170] Els van der Meijden, René W. A. Janssens, Chris Lauber, Jan Nico Bouwes Bavinck, Alexander E. Gorbalenya, et Mariet C. W. Feltkamp. Discovery of a New Human Polyomavirus Associated with Trichodysplasia Spinulosa in an Immunocompromized Patient. *PLoS Pathogens*, 6(7), July 2010.

ISSN 1553-7366. doi : 10.1371/journal.ppat.1001024. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2912394/.

[171] Simon Mazalrey, Dorian McIlroy, et Céline Bressollette-Bodin. BK polyomavirus : virus-cell interactions, host immune response, and viral pathogenesis. *Virologie*, 19(5) :8–24, September 2015. ISSN 1267-8694. doi : 10.1684/vir.2015.0624. URL http://www.jle.com/fr/revues/vir/e-docs/bk_polyomavirus_virus_cell_interactions_host_immune_response_and_viral_pathogenesis_305808/article.phtml?tab=texte.

[172] Simon Mazalrey. *Facteurs de pathogenèse au cours des infections à virus BK : polymorphisme génétique viral et réponse immunitaire antivirale*. These de doctorat, Nantes, October 2016. URL https://www.theses.fr/2016NANT1007.

[173] Christopher B. Buck, Koenraad Van Doorslaer, Alberto Peretti, Eileen M. Geoghegan, Michael J. Tisza, Ping An, Joshua P. Katz, James M. Pipas, Alison A. McBride, Alvin C. Camus, Alexa J. McDermott, Jennifer A. Dill, Eric Delwart, Terry F. F. Ng, Kata Farkas, Charlotte Austin, Simona Kraberger, William Davison, Diana V. Pastrana, et Arvind Varsani. The Ancient Evolutionary History of Polyomaviruses. *PLoS Pathogens*, 12(4), April 2016. ISSN 1553-7366. doi : 10.1371/journal.ppat.1005574. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4836724/.

[174] Elliot J Lefkowitz, Donald M Dempsey, Robert Curtis Hendrickson, Richard J Orton, Stuart G Siddell, et Donald B Smith. Virus taxonomy : the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, 46(Database issue) :D708–D717, January 2018. ISSN 0305-1048. doi : 10.1093/nar/gkx932. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753373/.

[175] Tobias Allander, Kalle Andreasson, Shawon Gupta, Annelie Bjerkner, Gordana Bogdanovic, Mats A. A. Persson, Tina Dalianis, Torbjörn Ramqvist, et Björn Andersson. Identification of a Third Human Polyomavirus. *Journal of Virology*, 81(8) :4130–4136, April 2007. ISSN 0022-538X. doi : 10.1128/JVI.00028-07. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1866148/.

[176] Anne M Gaynor, Michael D Nissen, David M Whiley, Ian M Mackay, Stephen B Lambert, Guang Wu, Daniel C Brennan, Gregory A Storch, Theo P Sloots, et David Wang. Identification of a Novel Polyomavirus from Patients with Acute Respiratory Tract Infections. *PLoS Pathogens*, 3(5), May 2007. ISSN 1553-7366. doi : 10.1371/journal.ppat.0030064. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1864993/.

[177] Rachel M. Schowalter, Diana V. Pastrana, Katherine A. Pumphrey, Adam L. Moyer, et Christopher B. Buck. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. *Cell Host & Microbe*, 7(6) :509–515, June 2010. ISSN 1934-6069. doi : 10.1016/j.chom.2010.05.006.

[178] Nelly Scuda, Jörg Hofmann, Sébastien Calvignac-Spencer, Klemens Ruprecht, Peter Liman, Joachim Kühn, Hartmut Hengel, et Bernhard Ehlers. A Novel Human Polyomavirus Closely Related to the African Green Monkey-Derived Lymphotropic Polyomavirus ? *Journal of Virology*, 85(9) :4586–4590, May 2011. ISSN 0022-538X. doi : 10.1128/JVI.02602-10. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3126223/.

[179] Erica A. Siebrasse, Alejandro Reyes, Efrem S. Lim, Guoyan Zhao, Rajhab S. Mkakosya, Mark J. Manary, Jeffrey I. Gordon, et David Wang. Identification of MW Polyomavirus, a Novel Polyomavirus in Human Stool. *Journal of Virology*, 86(19) :10321–10326, October 2012. ISSN 0022-538X. doi : 10.1128/JVI.01210-12. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3457274/.

[180] Efrem S. Lim, Alejandro Reyes, Martin Antonio, Debasish Saha, Usman N. Ikumapayi, Mitchell Adeyemi, O. Colin Stine, Rebecca Skelton, Daniel C. Brennan, Rajhab S. Mkakosya, Mark J. Manary, Jeffrey I. Gordon, et David Wang. Discovery of STL polyomavirus, a polyomavirus of ancestral recombinant origin that encodes a unique T antigen by alternative splicing. *Virology*, 436(2) : 295–303, February 2013. ISSN 1096-0341. doi : 10.1016/j.virol.2012.12.005.

[181] Sarah Korup, Janita Rietscher, Sébastien Calvignac-Spencer, Franziska Trusch, Jörg Hofmann, Ugo Moens, Igor Sauer, Sebastian Voigt, Rosa Schmuck, et Bernhard Ehlers. Identification of a Novel Human Polyomavirus in Organs of the Gastrointestinal Tract. 8(3), 2013. doi : 10.1371/journal.pone.0058021. Publisher : Robert Koch-Institut, Infektionskrankheiten / Erreger.

[182] Nischay Mishra, Marcus Pereira, Roy H. Rhodes, Ping An, James M. Pipas, Komal Jain, Amit Kapoor, Thomas Briese, Phyllis L. Faust, et W. Ian Lipkin. Identification of a Novel Polyomavirus in a Pancreatic Transplant Recipient With Retinal Blindness and Vasculitic Myopathy. *The Journal of Infectious Diseases*, 210(10) :1595–1599, November 2014. ISSN 0022-1899. doi : 10.1093/infdis/jiu250. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4334791/.

[183] Tarik Gheit, Sankhadeep Dutta, Javier Oliver, Alexis Robitaille, Shalaka Hampras, Jean-Damien Combes, Sandrine McKay-Chopin, Florence Le Calvez-Kelm, Neil Fenske, Basil Cherpelis, Anna R. Giuliano, Silvia Franceschi, James McKay, Dana E. Rollison, et Massimo Tommasino. Isolation and characterization of a novel putative human polyomavirus. *Virology*, 506 :45–54, 2017. ISSN 1096-0341. doi : 10.1016/j.virol.2017.03.007.

[184] Brian D. Ondov, Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, et Adam M. Phillippy. Mash Screen : high-throughput sequence containment estimation for genome discovery. *Genome Biology*, 20(1) :232, November 2019. ISSN 1474-760X. doi : 10.1186/s13059-019-1841-x. URL https://doi.org/10.1186/s13059-019-1841-x.

[185] R Hyde-DeRuyscher et G G Carmichael. Polyomavirus early-late switch is not regulated at the level of transcription initiation and is associated with changes in RNA processing. *Proceedings of the National Academy of Sciences of the United States of America*, 85(23) :8993–8997, December 1988. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC282648/.

[186] Friederike Neumann, Sophie Borchert, Claudia Schmidt, Rudolph Reimer, Heinrich Hohenberg, Nicole Fischer, et Adam Grundhoff. Replication, Gene Expression and Particle Production by a Consensus Merkel Cell Polyomavirus (MCPyV) Genome. *PLOS ONE*, 6(12) :e29112, December 2011. ISSN 1932-6203. doi : 10.1371/journal.pone.0029112. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0029112. Publisher : Public Library of Science.

[187] Sophie Borchert, Manja Czech-Sioli, Friederike Neumann, Claudia Schmidt, Peter Wimmer, Thomas Dobner, Adam Grundhoff, et Nicole Fischer. High-Affinity Rb Binding, p53 Inhibition, Subcellular Localization, and Transformation by Wild-Type or Tumor-Derived Shortened Merkel Cell

Polyomavirus Large T Antigens. *Journal of Virology*, 88(6) :3144–3160, March 2014. ISSN 0022-538X. doi : 10.1128/JVI.02916-13. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3957953/.

[188] Joseph J. Carter, Matthew D. Daugherty, Xiaojie Qi, Anjali Bheda-Malge, Gregory C. Wipf, Kristin Robinson, Ann Roman, Harmit S. Malik, et Denise A. Galloway. Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. *Proceedings of the National Academy of Sciences of the United States of America*, 110(31) :12744–12749, July 2013. ISSN 1091-6490. doi : 10.1073/pnas.1303526110.

[189] Camila Freze Baez, Rafael Brandão Varella, Sonia Villani, et Serena Delbue. Human Polyomaviruses : The Battle of Large and Small Tumor Antigens. *Virology : Research and Treatment*, 8, December 2017. ISSN 1178-122X. doi : 10.1177/1178122X17744785. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721967/.

[190] Hélène Laude. *Rôle du Polyomavirus de Merkel dans les carcinomes à cellules de Merkel*. phdthesis, Université René Descartes - Paris V, November 2012. URL https://tel.archives-ouvertes.fr/tel-00801219.

[191] Hana Španielová, Martin Fraiberk, Jiřina Suchanová, Jakub Soukup, et Jitka Forstová. The encapsidation of polyomavirus is not defined by a sequence-specific encapsidation signal. *Virology*, 450-451 :122–131, February 2014. ISSN 0042-6822. doi : 10.1016/j.virol.2013.12.010. URL http://www.sciencedirect.com/science/article/pii/S0042682213006739.

[192] Akira Nakanishi, Dorothy Shum, Hiroshi Morioka, Eiko Otsuka, et Harumi Kasamatsu. Interaction of the Vp3 nuclear localization signal with the importin alpha 2/beta heterodimer directs nuclear entry of infecting simian virus 40. *Journal of Virology*, 76(18) :9368–9377, September 2002. ISSN 0022-538X. doi : 10.1128/jvi.76.18.9368-9377.2002.

[193] Shauna M. Bennett, Nicole M. Broekema, et Michael J. Imperiale. BK polyomavirus : emerging pathogen. *Microbes and infection / Institut Pasteur*, 14(9) :672–683, August 2012. ISSN 1286-4579. doi : 10.1016/j.micinf.2012.02.002. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3568954/.

[194] Laura C. Ellis, Elizabeth Norton, Xin Dang, et Igor J. Koralnik. Agnogene Deletion in a Novel Pathogenic JC Virus Isolate Impairs VP1 Expression and Virion Production. *PLOS ONE*, 8(11) : e80840, November 2013. ISSN 1932-6203. doi : 10.1371/journal.pone.0080840. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0080840. Publisher : Public Library of Science.

[195] Mona Johannessen, Mari Walquist, Nancy Gerits, Marte Dragset, Anne Spang, et Ugo Moens. BKV Agnoprotein Interacts with α-Soluble N-Ethylmaleimide-Sensitive Fusion Attachment Protein, and Negatively Influences Transport of VSVG-EGFP. *PLOS ONE*, 6(9) :e24489, September 2011. ISSN 1932-6203. doi : 10.1371/journal.pone.0024489. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0024489. Publisher : Public Library of Science.

[196] Nancy Gerits et Ugo Moens. Agnoprotein of mammalian polyomaviruses. *Virology*, 432(2) :316–326, October 2012. ISSN 0042-6822. doi : 10.1016/j.virol.2012.05.024. URL http://www.sciencedirect.com/science/article/pii/S0042682212002759.

[197] Michael J. Imperiale. Polyomavirus miRNAs : The Beginning. *Current opinion in virology*, 0 : 29–32, August 2014. ISSN 1879-6257. doi : 10.1016/j.coviro.2014.03.012. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4149923/.

[198] Reimar Johne, Christopher B. Buck, Tobias Allander, Walter J. Atwood, Robert L. Garcea, Michael J. Imperiale, Eugene O. Major, Torbjorn Ramqvist, et Leonard C. Norkin. Taxonomical developments in the family Polyomaviridae. *Archives of virology*, 156(9), September 2011. ISSN 0304-8608. doi : 10.1007/s00705-011-1008-x. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3815707/.

[199] Andi Krumbholz, Olaf R. P. Bininda-Emonds, Peter Wutzler, et Roland Zell. Evolution of four BK virus subtypes. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 8(5) :632–643, September 2008. ISSN 1567-1348. doi : 10.1016/j.meegid.2008.05.006.

[200] Chunqing Luo, Marta Bueno, Jeffrey Kant, Jeremy Martinson, et Parmjeet Randhawa. Genotyping schemes for polyomavirus BK, using gene-specific phylogenetic trees and single nucleotide polymorphism analysis. *Journal of Virology*, 83(5) :2285–2297, March 2009. ISSN 1098-5514. doi : 10.1128/JVI.02180-08.

[201] Caiqin Hu, Ying Huang, Juwei Su, Mengyan Wang, Qihui Zhou, et Biao Zhu. The prevalence and isolated subtypes of BK polyomavirus reactivation among patients infected with human immunodeficiency virus-1 in southeastern China. *Archives of Virology*, 163(6) :1463–1468, 2018. ISSN 0304-8608. doi : 10.1007/s00705-018-3724-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5958166/.

[202] H F Wunderink, C S De Brouwer, L Gard, J W De Fijter, A C M Kroes, J I Rotmans, et M C W Feltkamp. Source and Relevance of the BK Polyomavirus Genotype for Infection After Kidney Transplantation. *Open Forum Infectious Diseases*, 6(3), February 2019. ISSN 2328-8957. doi : 10.1093/ofid/ofz078. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6440680/.

[203] Maria Chiara G. Monaco, Peter N. Jensen, Jean Hou, Linda C. Durham, et Eugene O. Major. Detection of JC Virus DNA in Human Tonsil Tissue : Evidence for Site of Initial Viral Infection. *Journal of Virology*, 72(12) :9918–9923, December 1998. ISSN 0022-538X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC110504/.

[204] Jane Kuypers, Angela P. Campbell, Katherine A. Guthrie, Nancy L. Wright, Janet A. Englund, Lawrence Corey, et Michael Boeckh. WU and KI Polyomaviruses in Respiratory Samples from Allogeneic Hematopoietic Cell Transplant Recipients. *Emerging Infectious Diseases*, 18(10) :1580–1588, October 2012. ISSN 1080-6040. doi : 10.3201/eid1810.120477. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3471632/.

[205] Raphael P. Viscidi, Dana E. Rollison, Vernon K. Sondak, Barbara Silver, Jane L. Messina, Anna R. Giuliano, William Fulp, Abidemi Ajidahun, et Daniela Rivanera. Age-Specific Seroprevalence of

Merkel Cell Polyomavirus, BK Virus, and JC Virus ? *Clinical and Vaccine Immunology : CVI*, 18 (10) :1737–1743, October 2011. ISSN 1556-6811. doi : 10.1128/CVI.05175-11. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3187023/.

[206] Annika Stolt, Kestutis Sasnauskas, Pentti Koskela, Matti Lehtinen, et Joakim Dillner. Seroepidemiology of the human polyomaviruses. *The Journal of General Virology*, 84(Pt 6) :1499–1504, June 2003. ISSN 0022-1317. doi : 10.1099/vir.0.18842-0.

[207] Bernard N Fields, David M Knipe, et Peter M Howley. *Fields virology*. Lippincott-Raven Publishers, Philadelphia, 1996. ISBN 978-0-7817-0253-9. OCLC : 32512536.

[208] Daisy Maria Machado, Maria Cristina Fink, Cláudio Sérgio Pannuti, Regina Célia de Menezes Succi, Alessandra Aparecida Machado, Fabiana Bononi do Carmo, Aída de Fátima Barbosa Gouvêa, Paulo Roberto Urbano, Suenia Vasconcelos Beltrão, Isabel Cristina Lopes dos Santos, et Clarisse Martins Machado. Human polyomaviruses JC and BK in the urine of Brazilian children and adolescents vertically infected by HIV. *Memórias do Instituto Oswaldo Cruz*, 106(8) :931–935, December 2011. ISSN 0074-0276. doi : 10.1590/S0074-02762011000800006. URL http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0074-02762011000800006&lng=en&nrm=iso&tlng=en.

[209] Jaime M. Kean, Suchitra Rao, Michael Wang, et Robert L. Garcea. Seroepidemiology of human polyomaviruses. *PLoS pathogens*, 5(3) :e1000363, March 2009. ISSN 1553-7374. doi : 10.1371/journal.ppat.1000363.

[210] Hans H. Hirsch, Cinthia B. Drachenberg, Juerg Steiger, et Emilo Ramos. *Polyomavirus-associated Nephropathy in Renal Transplantation : Critical Issues of Screening and Management*. Landes Bioscience, 2013. URL https://www.ncbi.nlm.nih.gov/books/NBK6388/. Publication Title : Madame Curie Bioscience Database [Internet].

[211] Deepak K. Rajpoot, Allan Gomez, Walter Tsang, et Allan Shanberg. Ureteric and urethral stenosis : a complication of BK virus infection in a pediatric renal transplant patient. *Pediatric Transplantation*, 11(4) :433–435, June 2007. ISSN 1397-3142. doi : 10.1111/j.1399-3046.2006.00673.x.

[212] Hannah Imlay, Hu Xie, Wendy M. Leisenring, Elizabeth R. Duke, Louise E. Kimball, Meei-Li Huang, Steven A. Pergam, Joshua A. Hill, Keith R. Jerome, Filippo Milano, W. Garrett Nichols, Phillip S. Pang, Hans H. Hirsch, Ajit P. Limaye, et Michael Boeckh. Presentation of BK polyomavirus–associated hemorrhagic cystitis after allogeneic hematopoietic cell transplantation. *Blood Advances*, 4(4) :617–628, February 2020. ISSN 2473-9529. doi : 10.1182/bloodadvances.2019000802. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7042995/.

[213] Mengxi Jiang, Johanna R. Abend, Silas F. Johnson, et Michael J. Imperiale. The Role of Polyomaviruses in Human Disease. *Virology*, 384(2) :266–273, February 2009. ISSN 0042-6822. doi : 10.1016/j.virol.2008.09.027. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2661150/.

[214] D. V. Coleman, E. F. Mackenzie, S. D. Gardner, J. M. Poulding, B. Amer, et W. J. Russell. Human polyomavirus (BK) infection and ureteric stenosis in renal allograft recipients. *Journal of Clinical Pathology*, 31(4) :338–347, April 1978. ISSN 0021-9746. doi : 10.1136/jcp.31.4.338.

[215] G Gargiulo, L Orlando, F Alberani, G Crabu, A Di Maio, L Duranti, A Errico, S Liptrott, R Pitrone, S Santarone, C Soliman, A Trunfio, C Selleri, B Bruno, S Mammoliti, et F Pane. Haemorrhagic cystitis in haematopoietic stem cell transplantation (HSCT) : a prospective observational study of incidence and management in HSCT centres within the GITMO network (Gruppo Italiano Trapianto Midollo Osseo). *ecancermedicalscience*, 8, April 2014. ISSN 1754-6605. doi : 10.3332/ecancer.2014.420. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998658/.

[216] de Vries Catherine R. et Freiha Fuad S. Hemorrhagic Cystitis : A Review. *Journal of Urology*, 143(1) :1–9, January 1990. doi : 10.1016/S0022-5347(17)39848-8. URL https://www.auajournals.org/doi/abs/10.1016/S0022-5347%2817%2939848-8. Publisher : WoltersKluwer.

[217] LK Dropulic et RJ Jones. Polyomavirus BK infection in blood and marrow transplant recipients. *Bone marrow transplantation*, 41(1) :11–18, January 2008. ISSN 0268-3369. doi : 10.1038/sj.bmt.1705886. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066131/.

[218] Veronique Erard, Hyung Woo Kim, Lawrence Corey, Ajit Limaye, Meei-Li Huang, David Myerson, Chris Davis, et Michael Boeckh. BK DNA viral load in plasma : evidence for an association with hemorrhagic cystitis in allogeneic hematopoietic cell transplant recipients. *Blood*, 106(3) :1130–1132, August 2005. ISSN 0006-4971. doi : 10.1182/blood-2004-12-4988.

[219] Jérémie Corneille et David Boutolleau. Infections à virus BK après allogreffe de cellules souches hématopoïétiques. *Virologie*, 15(2) :115–125, May 2011. ISSN 1267-8694. doi : 10.1684/vir.2011.0402. URL http://www.jle.com/fr/revues/vir/e-docs/infections_a_virus_bk_apres_allogreffe_de_cellules_souches_hematopoietiques_288625/article.phtml?tab=texte.

[220] Corinna Schmitt, Lubna Raggub, Silvia Linnenweber-Held, Ortwin Adams, Anke Schwarz, et Albert Heim. Donor origin of BKV replication after kidney transplantation. *Journal of Clinical Virology : The Official Publication of the Pan American Society for Clinical Virology*, 59(2) :120–125, February 2014. ISSN 1873-5967. doi : 10.1016/j.jcv.2013.11.009.

[221] V. Nickeleit, H. H. Hirsch, I. F. Binet, F. Gudat, O. Prince, P. Dalquen, G. Thiel, et M. J. Mihatsch. Polyomavirus infection of renal allograft recipients : from latent infection to manifest disease. *Journal of the American Society of Nephrology : JASN*, 10(5) :1080–1089, May 1999. ISSN 1046-6673.

[222] Hans H. Hirsch, Daniel C. Brennan, Cinthia B. Drachenberg, Fabrizio Ginevri, Jennifer Gordon, Ajit P. Limaye, Michael J. Mihatsch, Volker Nickeleit, Emilio Ramos, Parmjeet Randhawa, Ron Shapiro, Juerg Steiger, Manikkam Suthanthiran, et Jennifer Trofe. Polyomavirus-associated nephropathy in renal transplantation : interdisciplinary analyses and recommendations. *Transplantation*, 79(10) :1277–1286, May 2005. ISSN 0041-1337. doi : 10.1097/01.tp.0000156165.83160.09.

[223] Jeffrey Pina, Robert Jordan Ontiveros, Niroshika Keppetipola, et Nikolas Nikolaidis. A Bioinformatics Approach to Discover the Evolutionary Origin of the PTBP Splicing Regulators. *The FASEB Journal*, 32(1_supplement) :802.16–802.16, April 2018. ISSN 0892-6638. doi : 10.1096/fasebj.2018.32.1_supplement.802.16. URL https://www.fasebj.org/doi/abs/10.1096/fasebj.2018.32.1_supplement.802.16. Publisher : Federation of American Societies for Experimental Biology.

[224] Fiona Robinson, Richard J. Jackson, et Christopher W. J. Smith. Expression of Human nPTB Is Limited by Extreme Suboptimal Codon Content. *PLOS ONE*, 3(3) :e1801, March 2008. ISSN 1932-6203. doi : 10.1371/journal.pone.0001801. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0001801. Publisher : Public Library of Science.

[225] J. R. Lobry et C. Gautier. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acids Research*, 22 (15) :3174–3180, August 1994. ISSN 0305-1048.

[226] Jack Kyte et Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1) :105–132, May 1982. ISSN 0022-2836. doi : 10.1016/0022-2836(82)90515-0. URL http://www.sciencedirect.com/science/article/pii/0022283682905150.

[227] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 46(D1) :D8–D13, 2018. ISSN 1362-4962. doi : 10.1093/nar/gkx1095.

[228] Peter J. Huber. Robust Statistics. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi : 10.1007/978-3-642-04898-2_594. URL https://doi.org/10.1007/978-3-642-04898-2_594.

[229] LaDeana W. Hillier, Webb Miller, Ewan Birney, Wesley Warren, Ross C. Hardison, Chris P. Ponting, Peer Bork, David W. Burt, Martien A. M. Groenen, Mary E. Delany, Jerry B. Dodgson, Asif T. Chinwalla, Paul F. Cliften, Sandra W. Clifton, Kimberly D. Delehaunty, Catrina Fronick, Robert S. Fulton, Tina A. Graves, Colin Kremitzki, Dan Layman, Vincent Magrini, John D. McPherson, Tracie L. Miner, Patrick Minx, William E. Nash, Michael N. Nhan, Joanne O. Nelson, Lachlan G. Oddy, Craig S. Pohl, Jennifer Randall-Maher, Scott M. Smith, John W. Wallis, Shiaw-Pyng Yang, Michael N. Romanov, Catherine M. Rondelli, Bob Paton, Jacqueline Smith, David Morrice, Laura Daniels, Helen G. Tempest, Lindsay Robertson, Julio S. Masabanda, Darren K. Griffin, Alain Vignal, Valerie Fillon, Lina Jacobbson, Susanne Kerje, Leif Andersson, Richard P. M. Crooijmans, Jan Aerts, Jan J. van der Poel, Hans Ellegren, Randolph B. Caldwell, Simon J. Hubbard, Darren V. Grafham, Andrzej M. Kierzek, Stuart R. McLaren, Ian M. Overton, Hiroshi Arakawa, Kevin J. Beattie, Yuri Bezzubov, Paul E. Boardman, James K. Bonfield, Michael D. R. Croning, Robert M. Davies, Matthew D. Francis, Sean J. Humphray, Carol E. Scott, Ruth G. Taylor, Cheryll Tickle, William R. A. Brown, Jane Rogers, Jean-Marie Buerstedde, Stuart A. Wilson, Lisa Stubbs, Ivan Ovcharenko, Laurie Gordon, Susan Lucas, Marcia M. Miller, Hidetoshi Inoko, Takashi Shiina, Jim Kaufman, Jan Salomonsen, Karsten Skjoedt, Gane Ka-Shu Wong, Jun Wang, Bin Liu, Jian Wang, Jun Yu, Huanming Yang, Mikhail Nefedov, Maxim Koriabine, Pieter J. deJong, Leo Goodstadt, Caleb Webber, Nicholas J. Dickens, Ivica Letunic, Mikita Suyama, David Torrents, Christian von Mering, Evgeny M. Zdobnov, Kateryna Makova, Anton Nekrutenko, Laura Elnitski, Pallavi Eswara, David C. King, Shan Yang, Svitlana Tyekucheva, Anusha Radakrishnan, Robert S. Harris, Francesca Chiaromonte, James Taylor, Jianbin He, Monique Rijnkels, Sam Griffiths-Jones, Abel Ureta-Vidal, Michael M. Hoffman, Jessica Severin, Stephen M. J. Searle, Andy S. Law, David Speed, Dave Waddington, Ze Cheng, Eray Tuzun, Evan Eichler, Zhirong Bao, Paul Flicek, David D.

Shteynberg, Michael R. Brent, Jacqueline M. Bye, Elizabeth J. Huckle, Sourav Chatterji, Colin Dewey, Lior Pachter, Andrei Kouranov, Zissimos Mourelatos, Artemis G. Hatzigeorgiou, Andrew H. Paterson, Robert Ivarie, Mikael Brandstrom, Erik Axelsson, Niclas Backstrom, Sofia Berlin, Matthew T. Webster, Olivier Pourquie, Alexandre Reymond, Catherine Ucla, Stylianos E. Antonarakis, Manyuan Long, J. J. Emerson, Esther Betrán, Isabelle Dupanloup, Henrik Kaessmann, Angie S. Hinrichs, Gill Bejerano, Terrence S. Furey, Rachel A. Harte, Brian Raney, Adam Siepel, W. James Kent, David Haussler, Eduardo Eyras, Robert Castelo, Josep F. Abril, Sergi Castellano, Francisco Camara, Genis Parra, Roderic Guigo, Guillaume Bourque, Glenn Tesler, Pavel A. Pevzner, Arian Smit, Lucinda A. Fulton, Elaine R. Mardis, Richard K. Wilson, International Chicken Genome Sequencing Consortium, Overall coordination :, sequence and assembly : Genome fingerprint map, Mapping :, cDNA sequencing :, Other sequencing and libraries :, Analysis and annotation :, et Project management :. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018) :695–716, December 2004. ISSN 1476-4687. doi : 10.1038/nature03154. URL https://www.nature.com/articles/nature03154. Number : 7018 Publisher : Nature Publishing Group.

[230] Erik L. L. Sonnhammer et Eugene V. Koonin. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in genetics : TIG*, 18(12) :619–620, December 2002. ISSN 0168-9525. doi : 10.1016/s0168-9525(02)02793-2.

[231] Shelley D. Copley. Evolution of new enzymes by gene duplication and divergence. *The FEBS journal*, 287(7) :1262–1283, April 2020. ISSN 1742-4658. doi : 10.1111/febs.15299.

[232] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, et Takashi Miyata. MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30 (14) :3059–3066, July 2002. ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC135756/.

[233] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4) :540–552, April 2000. ISSN 0737-4038. doi : 10.1093/oxfordjournals.molbev.a026334.

[234] Alexandros Stamatakis. RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9) :1312–1313, May 2014. ISSN 1367-4811. doi : 10.1093/bioinformatics/btu033.

[235] Sudhir Kumar, Glen Stecher, Michael Suleski, et S. Blair Hedges. TimeTree : A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, 34(7) :1812–1819, 2017. ISSN 1537-1719. doi : 10.1093/molbev/msx116.

[236] D. F. Robinson et L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1) :131–147, February 1981. ISSN 0025-5564. doi : 10.1016/0025-5564(81)90043-2. URL http://www.sciencedirect.com/science/article/pii/0025556481900432.

[237] Víctor Soria-Carrasco, Gerard Talavera, Javier Igea, et Jose Castresana. The K tree score : quantification of differences in the relative branch length and topology of phylogenetic trees. *Bioinformatics (Oxford, England)*, 23(21) :2954–2956, November 2007. ISSN 1367-4811. doi : 10.1093/bioinformatics/btm466.

[238] Andrew D. Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Andrey G. Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, José Carlos Marugán, Carla Cummins, Claire Davidson, Kamalkumar Dodiya, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, Thomas Maurel, Mark McDowall, Aoife McMahon, Shamika Mohanan, Benjamin Moore, Michael Nuhn, Denye N. Oheh, Anne Parker, Andrew Parton, Mateus Patricio, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M. Schmitt, Helen Schuilenburg, Dan Sheppard, Mira Sycheva, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Bethany Flint, Adam Frankish, Sarah E. Hunt, Garth IIsley, Myrto Kostadima, Nick Langridge, Jane E. Loveland, Fergal J. Martin, Joannella Morales, Jonathan M. Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, Stephen J. Trevanion, Fiona Cunningham, Kevin L. Howe, Daniel R. Zerbino, et Paul Flicek. Ensembl 2020. *Nucleic Acids Research*, 48(D1) :D682–D688, January 2020. ISSN 0305-1048. doi : 10.1093/nar/gkz966. URL https://academic.oup.com/nar/article/48/D1/D682/5613682. Publisher : Oxford Academic.

[239] Celine Scornavacca, Khalid Belkhir, Jimmy Lopez, Rémy Dernat, Frédéric Delsuc, Emmanuel J. P. Douzery, et Vincent Ranwez. OrthoMaM v10 : Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. *Molecular Biology and Evolution*, 36(4) :861–862, April 2019. ISSN 0737-4038. doi : 10.1093/molbev/msz015. URL https://academic.oup.com/mbe/article/36/4/861/5303840. Publisher : Oxford Academic.

[240] T. Caspersson, S. Farber, G. E. Foley, J. Kudynowski, E. J. Modest, E. Simonsson, U. Wagh, et L. Zech. Chemical differentiation along metaphase chromosomes. *Experimental Cell Research*, 49 (1) :219–222, January 1968. ISSN 0014-4827. doi : 10.1016/0014-4827(68)90538-7.

[241] Gerald P. Holmquist. Evolution of chromosome bands : Molecular ecology of noncoding DNA. *Journal of Molecular Evolution*, 28(6) :469–486, June 1989. ISSN 1432-1432. doi : 10.1007/BF02602928. URL https://doi.org/10.1007/BF02602928.

[242] Sylvain Glémin, Peter F. Arndt, Philipp W. Messer, Dmitri Petrov, Nicolas Galtier, et Laurent Duret. Quantification of GC-biased gene conversion in the human genome. *Genome Research*, 25(8) : 1215–1228, January 2015. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.185488.114. URL http://genome.cshlp.org/content/25/8/1215. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab.

[243] Ruth Hershberg et Dmitri A. Petrov. General rules for optimal codon choice. *PLoS genetics*, 5(7) : e1000556, July 2009. ISSN 1553-7404. doi : 10.1371/journal.pgen.1000556.

[244] Christina E. Brule et Elizabeth J. Grayhack. Synonymous Codons : Choose Wisely for Expression. *Trends in genetics : TIG*, 33(4) :283–297, 2017. ISSN 0168-9525. doi : 10.1016/j.tig.2017.02.001.

[245] Zachary R. Newman, Janet M. Young, Nicholas T. Ingolia, et Gregory M. Barton. Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. *Proceedings*

of the National Academy of Sciences of the United States of America, 113(10) :E1362–1371, March 2016. ISSN 1091-6490. doi : 10.1073/pnas.1518976113.

[246] Benjamin L. Lampson, Nicole L. K. Pershing, Joseph A. Prinz, Joshua R. Lacsina, William F. Marzluff, Christopher V. Nicchitta, David M. MacAlpine, et Christopher M. Counter. Rare codons regulate KRas oncogenesis. *Current biology : CB*, 23(1) :70–75, January 2013. ISSN 1879-0445. doi : 10.1016/j.cub.2012.11.031.

[247] Jingjing Fu, Yunkun Dang, Christopher Counter, et Yi Liu. Codon usage regulates human KRAS expression at both transcriptional and translational levels. *The Journal of Biological Chemistry*, 293(46) :17929–17940, 2018. ISSN 1083-351X. doi : 10.1074/jbc.RA118.004908.

[248] Niroshika Keppetipola, Shalini Sharma, Qin Li, et Douglas L. Black. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. *Critical Reviews in Biochemistry and Molecular Biology*, 47(4) :360–378, August 2012. ISSN 1549-7798. doi : 10.3109/10409238.2012.691456.

[249] S. D. Ferris et G. S. Whitt. Evolution of the differential regulation of duplicate genes after polyploidization. *Journal of Molecular Evolution*, 12(4) :267–317, April 1979. ISSN 0022-2844. doi : 10.1007/BF01732026.

[250] Aldo Donizetti, Marcella Fiengo, Sergio Minucci, et Francesco Aniello. Duplicated zebrafish relaxin-3 gene shows a different expression pattern from that of the co-orthologue gene. *Development, Growth & Differentiation*, 51(8) :715–722, October 2009. ISSN 1440-169X. doi : 10.1111/j.1440-169X.2009.01131.x.

[251] Katerina Guschanski, Maria Warnefors, et Henrik Kaessmann. The evolution of duplicate gene expression in mammalian organs. *Genome Research*, 27(9) :1461–1474, 2017. ISSN 1549-5469. doi : 10.1101/gr.215566.116.

[252] Shiri Freilich, Tim Massingham, Eric Blanc, Leon Goldovsky, et Janet M. Thornton. Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome Biology*, 7(10) :R89, 2006. ISSN 1474-760X. doi : 10.1186/gb-2006-7-10-r89.

[253] Keith L. Adams, Richard Cronn, Ryan Percifield, et Jonathan F. Wendel. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8) :4649–4654, April 2003. ISSN 0027-8424. doi : 10.1073/pnas.0630618100.

[254] Eugene V. Makeyev, Jiangwen Zhang, Monica A. Carrasco, et Tom Maniatis. The MicroRNA miR-124 Promotes Neuronal Differentiation by Triggering Brain-Specific Alternative Pre-mRNA Splicing. *Molecular cell*, 27(3) :435–448, August 2007. ISSN 1097-2765. doi : 10.1016/j.molcel.2007.07.015. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3139456/.

[255] John K. Vuong, Chia-Ho Lin, Min Zhang, Liang Chen, Douglas L. Black, et Sika Zheng. PTBP1 and PTBP2 Serve Both Specific and Redundant Functions in Neuronal Pre-mRNA Splicing. *Cell Reports*, 17(10) :2766–2775, 2016. ISSN 2211-1247. doi : 10.1016/j.celrep.2016.11.034.

[256] Simon Laurin-Lemay, Nicolas Rodrigue, Nicolas Lartillot, et Hervé Philippe. Conditional Approximate Bayesian Computation : A New Approach for Across-Site Dependency in High-Dimensional Mutation-Selection Models. *Molecular Biology and Evolution*, 35(11) :2819–2834, 2018. ISSN 1537-1719. doi : 10.1093/molbev/msy173.

[257] Eva Maria Novoa, Irwin Jungreis, Olivier Jaillon, et Manolis Kellis. Elucidation of Codon Usage Signatures across the Domains of Life. *Molecular Biology and Evolution*, 36(10) :2328–2339, 2019. ISSN 1537-1719. doi : 10.1093/molbev/msz124.

[258] D. S. Peabody. Translation initiation at non-AUG triplets in mammalian cells. *The Journal of Biological Chemistry*, 264(9) :5031–5035, March 1989. ISSN 0021-9258.

[259] Ivaylo P. Ivanov, Andrew E. Firth, Audrey M. Michel, John F. Atkins, et Pavel V. Baranov. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Research*, 39(10) :4220–4234, May 2011. ISSN 1362-4962. doi : 10.1093/nar/gkr007.

[260] Kazutaka Katoh et Daron M. Standley. MAFFT Multiple Sequence Alignment Software Version 7 : Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4) :772–780, January 2013. ISSN 0737-4038, 1537-1719. doi : 10.1093/molbev/mst010. URL http://mbe.oxfordjournals.org/content/30/4/772.

[261] Sara Goodwin, John D. McPherson, et W. Richard McCombie. Coming of age : ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6) :333–351, June 2016. ISSN 1471-0064. doi : 10.1038/nrg.2016.49. URL https://www.nature.com/articles/nrg.2016.49. Number : 6 Publisher : Nature Publishing Group.

[262] F. Sanger, S. Nicklen, et A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12) :5463–5467, December 1977. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/.

[263] J. F. Hess, T. A. Kohl, M. Kotrová, K. Rönsch, T. Paprotka, V. Mohr, T. Hutzenlaub, M. Brüggemann, R. Zengerle, S. Niemann, et N. Paust. Library preparation for next generation sequencing : A review of automation strategies. *Biotechnology Advances*, 41 :107537, July 2020. ISSN 0734-9750. doi : 10.1016/j.biotechadv.2020.107537. URL http://www.sciencedirect.com/science/article/pii/S0734975020300343.

[264] Shahid Raza et Ayesha Ameen. Nano pore Sequencing Technology : A Review. *International Journal of Advances in Scientific Research*, 3 :90, August 2017. doi : 10.7439/ijasr.v3i8.4333.

[265] Michael L. Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11 (1) :31–46, January 2010. ISSN 1471-0064. doi : 10.1038/nrg2626. URL https://www.nature.com/articles/nrg2626. Number : 1 Publisher : Nature Publishing Group.

[266] Jason R. Miller, Sergey Koren, et Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6) :315–327, June 2010. ISSN 0888-7543. doi : 10.1016/j.ygeno.2010.03.001. URL http://www.sciencedirect.com/science/article/pii/S0888754310000492.

[267] Knut Reinert, Ben Langmead, David Weese, et Dirk J. Evers. Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics*, 16(1) :133–151, 2015. doi : 10.1146/annurev-genom-090413-025358. URL https://doi.org/10.1146/annurev-genom-090413-025358. _eprint : https ://doi.org/10.1146/annurev-genom-090413-025358.

[268] Suying Bao, Rui Jiang, WingKeung Kwan, Binbin Wang, Xu Ma, et You-Qiang Song. Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics*, 56 : 406–414, April 2011. doi : 10.1038/jhg.2011.43.

[269] Jonathan M. Rothberg et John H. Leamon. The development and impact of 454 sequencing. *Nature Biotechnology*, 26(10) :1117–1124, October 2008. ISSN 1546-1696. doi : 10.1038/nbt1485. URL https://www.nature.com/articles/nbt1485. Number : 10 Publisher : Nature Publishing Group.

[270] Simon Ardui, Adam Ameur, Joris R. Vermeesch, et Matthew S. Hestand. Single molecule real-time (SMRT) sequencing comes of age : applications and utilities for medical diagnostics. *Nucleic Acids Research*, 46(5) :2159–2168, March 2018. ISSN 0305-1048. doi : 10.1093/nar/gky066. URL https://academic.oup.com/nar/article/46/5/2159/4833218. Publisher : Oxford Academic.

[271] Hengyun Lu, Francesca Giordano, et Zemin Ning. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, 14(5) :265–279, October 2016. ISSN 1672-0229. doi : 10.1016/j.gpb.2016.05.004. URL http://www.sciencedirect.com/science/article/pii/S1672022916301309.

[272] Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, et Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10) :1155–1162, October 2019. ISSN 1546-1696. doi : 10.1038/s41587-019-0217-9. URL https://www.nature.com/articles/s41587-019-0217-9. Number : 10 Publisher : Nature Publishing Group.

[273] Eric L. Delwart. Viral metagenomics. *Reviews in Medical Virology*, 17(2) :115–131, April 2007. ISSN 1052-9276. doi : 10.1002/rmv.532.

[274] Sander van Boheemen, Anneloes L. van Rijn, Nikos Pappas, Ellen C. Carbo, Ruben H. P. Vorderman, Igor Sidorov, Peter J. van 't Hof, Hailiang Mei, Eric C. J. Claas, Aloys C. M. Kroes, et Jutte J. C. de Vries. Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients. *The Journal of Molecular Diagnostics*, 22(2) :196–207, February 2020. ISSN 1525-1578. doi : 10.1016/j.jmoldx.2019.10.007. URL http://www.sciencedirect.com/science/article/pii/S1525157819304325.

[275] Timofey Skvortsov, Colin de Leeuwe, John P. Quinn, John W. McGrath, Christopher C. R. Allen, Yvonne McElarney, Catherine Watson, Ksenia Arkhipova, Rob Lavigne, et Leonid A. Kulakov. Metagenomic Characterisation of the Viral Community of Lough Neagh, the Largest Freshwater Lake in Ireland. *PLoS ONE*, 11(2), February 2016. ISSN 1932-6203. doi : 10.1371/journal.pone.0150361. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4771703/.

[276] Sven Warris, Elio Schijlen, Henri van de Geest, Rahulsimham Vegesna, Thamara Hesselink, Bas te Lintel Hekkert, Gabino Sanchez Perez, Paul Medvedev, Kateryna D. Makova, et Dick de Ridder. Correcting palindromes in long reads after whole-genome amplification. *BMC Genomics*, 19(1) : 798, November 2018. ISSN 1471-2164. doi : 10.1186/s12864-018-5164-1. URL https://doi.org/10.1186/s12864-018-5164-1.

[277] Venkatesh Kumar, Thomas Vollbrecht, Mark Chernyshev, Sanjay Mohan, Brian Hanst, Nicholas Bavafa, Antonia Lorenzo, Nikesh Kumar, Robert Ketteringham, Kemal Eren, Michael Golden, Michelli F Oliveira, et Ben Murrell. Long-read amplicon denoising. *Nucleic Acids Research*, 47(18) :e104, October 2019. ISSN 0305-1048. doi : 10.1093/nar/gkz657. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6765106/.

[278] S. Jane Flint, V. R Racaniello, Glenn F Rall, Anna Marie Skalka, et L. W Enquist. *Principles of virology*. 2015. ISBN 978-1-55581-933-0 978-1-55581-934-7 978-1-55581-951-4 978-1-55581-952-1 978-1-55581-895-1. OCLC : 914445879.

[279] Jasmina Vasilijevic, Noelia Zamarreño, Juan Carlos Oliveros, Ariel Rodriguez-Frandsen, Guillermo Gómez, Guadalupe Rodriguez, Mercedes Pérez-Ruiz, Sonia Rey, Isabel Barba, Francisco Pozo, Inmaculada Casas, Amelia Nieto, et Ana Falcón. Reduced accumulation of defective viral genomes contributes to severe outcome in influenza virus infected patients. *PLOS Pathogens*, 13(10) : e1006650, October 2017. ISSN 1553-7374. doi : 10.1371/journal.ppat.1006650. URL https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1006650. Publisher : Public Library of Science.

[280] Meghan Diefenbacher, Jiayi Sun, et Christopher B. Brooke. The parts are greater than the whole : the role of semi-infectious particles in influenza A virus biology. *Current Opinion in Virology*, 33 : 42–46, 2018. ISSN 1879-6265. doi : 10.1016/j.coviro.2018.07.002.

[281] Souichi Nukuzuma, Tomokazu Takasaka, Huai-Ying Zheng, Shan Zhong, Qin Chen, Tadaichi Kitamura, et Yoshiaki Yogo. Subtype I BK polyomavirus strains grow more efficiently in human renal epithelial cells than subtype IV strains. *The Journal of General Virology*, 87(Pt 7) :1893–1901, July 2006. ISSN 0022-1317. doi : 10.1099/vir.0.81698-0.

[282] L. Jin, P. E. Gibson, J. C. Booth, et J. P. Clewley. Genomic typing of BK virus in clinical specimens by direct sequencing of polymerase chain reaction products. *Journal of Medical Virology*, 41(1) : 11–17, September 1993. ISSN 0146-6615. doi : 10.1002/jmv.1890410104.

[283] Virginie Morel, Elodie Martin, Catherine François, François Helle, Justine Faucher, Thomas Mourez, Gabriel Choukroun, Gilles Duverlie, Sandrine Castelain, et Etienne Brochot. A Simple and Reliable Strategy for BK Virus Subtyping and Subgrouping. *Journal of Clinical Microbiology*, 55 (4) :1177–1185, 2017. ISSN 1098-660X. doi : 10.1128/JCM.01180-16.

[284] Dorian McIlroy, Franck Halary, et Céline Bressollette-Bodin. Intra-patient viral evolution in polyomavirus-related diseases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 374(1773) :20180301, 2019. ISSN 1471-2970. doi : 10.1098/rstb.2018.0301.

[285] Dorian McIlroy, Mario Hönemann, Ngoc-Khanh Nguyen, Paul Barbier, Cécile Peltier, Audrey Rodallec, Franck Halary, Emilie Przyrowski, Uwe Liebert, Maryvonne Hourmant, et Céline Bressollette-Bodin. Persistent BK Polyomavirus Viruria is Associated with Accumulation of VP1 Mutations and Neutralization Escape. *Viruses*, 12(8), 2020. ISSN 1999-4915. doi : 10.3390/v12080824.

[286] Pilar Domingo-Calap, Benjamin Schubert, Mélanie Joly, Morgane Solis, Meiggie Untrau, Raphael Carapito, Philippe Georgel, Sophie Caillard, Samira Fafi-Kremer, Nicodème Paul, Oliver Kohlbacher, Fernando González-Candelas, et Seiamak Bahram. An unusually high substitution rate in transplant-associated BK polyomavirus in vivo is further concentrated in HLA-C-bound viral peptides. *PLoS pathogens*, 14(10) :e1007368, 2018. ISSN 1553-7374. doi : 10.1371/journal.ppat.1007368.

[287] Marie Sémon, Jean R. Lobry, et Laurent Duret. No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Molecular Biology and Evolution*, 23(3) :523–529, March 2006. ISSN 0737-4038. doi : 10.1093/molbev/msj053.

[288] Xavier Hernandez-Alias, Hannah Benisty, Martin H. Schaefer, et Luis Serrano. Translational efficiency across healthy and tumor tissues is proliferation-related. *Molecular Systems Biology*, 16(3) : e9275, 2020. ISSN 1744-4292. doi : 10.15252/msb.20199275.

[289] Dongxue Wang, Basak Eraslan, Thomas Wieland, Björn Hallström, Thomas Hopf, Daniel Paul Zolg, Jana Zecha, Anna Asplund, Li-hua Li, Chen Meng, Martin Frejno, Tobias Schmidt, Karsten Schnatbaum, Mathias Wilhelm, Frederik Ponten, Mathias Uhlen, Julien Gagneur, Hannes Hahne, et Bernhard Kuster. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology*, 15(2) :e8503, February 2019. ISSN 1744-4292. doi : 10.15252/msb.20188503. URL https://www.embopress.org/doi/full/10.15252/msb.20188503. Publisher : John Wiley & Sons, Ltd.

[290] S. L. Chen, Y. P. Tsao, J. W. Lee, W. C. Sheu, et Y. T. Liu. Characterization and analysis of human papillomaviruses of skin warts. *Archives of Dermatological Research*, 285(8) :460–465, 1993. ISSN 0340-3696.

REMERCIEMENTS

Je remercie infiniment mes deux directeurs de thèse Ignacio (Nacho) G. Bravo et Samuel Alizon pour leur engagement dans ce doctorat. Un immense merci, Nacho, pour tout ce que tu as pu faire pour moi pendant le déroulement de ce doctorat. Quelle aventure ! Merci pour toute la confiance que tu as pu m'accorder, ainsi que pour la grande liberté que tu m'as donnée pendant ces trois années de doctorat. Pendant tout son déroulement, tu m'as guidé aussi bien scientifiquement que spirituellement, et même si nous affichions parfois tous deux un désaccord sur certains points (c'est ça la science !), j'ai adoré chacune de nos rencontres. Je n'aurais jamais pu terminer ce doctorat sans ton aide précieuse quant à la direction de certains travaux ainsi que sur la rédaction et la correction d'un nombre incalculable de rapports, posters et publications (et j'en passe). Merci pour tes conseils avisés, notamment sur tout ce qui concerne la vie scientifique et les mécanismes humains qui y sont associés. Et puis, pour ta gentillesse et pour tous les bons moments que nous avons pu passer ensemble. Grâce à toi, j'ai pu évoluer en tant que scientifique et je me sens prêt à réaliser plein de belles choses. Merci d'avoir cru en moi ! *Numquam dedite* ! Samuel, Jamais je n'aurais pu imaginer vivre une telle histoire après avoir virtuellement tapé à ta porte pour te demander un stage de master. Au final, c'est un peu grâce à toi que j'ai pu faire ce doctorat ! Merci pour tout Samuel ! Nous ne nous sommes pas beaucoup vus lors de cette thèse et j'en suis désolé. J'aurais voulu poursuivre nos travaux sur l'étude des papillomavirus et polyomavirus humains, et venir plus souvent sur la base Kayak. Mais ça ne t'a pas empêché d'être toujours là pour un conseil scientifique, pour des corrections avisées et pour un avis toujours pointu qui savait toucher sa cible. Grâce à toi, à ta rigueur professionnelle et à tes talents de communication scientifique, j'ai pu apprendre beaucoup de choses. Merci pour toute ton aide, du début à la fin de ce doctorat !

Je remercie tous les membres de l'UMR MIVEGEC, et plus précisément des équipes VIROSTYLE et ETE, qui forment un laboratoire détonant de créativité, de sympathie et de rencontres aussi improbables que géniales. Ce laboratoire a été un lieu formidable de rencontres, de discussions animées et de moments forts. Je remercie plus particulièrement mes copains et collègues proches Fiona, Fanni, Josquin, Cécile, Vanina, Alba, Laura, Marion, Rémy, Gonché, Mircea, Yannis, Soraya, Luce, Claire, Massilva et Arthur. Mais je souhaite aussi remercier avec tout autant d'entrain toutes les personnes qui ne sont pas ici citées, avec qui j'ai pu entretenir de longues et passionnantes discussions. Décrire tout ce que vous m'avez apporté ces dernières années serait bien trop long (et c'est tant mieux !). Merci à vous tous pour votre sympathie et les bons moments que nous avons passés ensemble.

Je tiens à remercier toutes les personnes qui ont suivi le déroulement de ce doctorat, en tant que membres de comité ou de mon jury. Vos conseils et critiques avisés m'ont permis de prendre du recul sur mes travaux, de me motiver et m'ont fait considérablement avancer dans mes

recherches. Je tiens particulièrement à remercier : Celine Scornavacca, Michel Segondy, Frédéric Delsuc, Céline Bressollette, Anna-Sophie Fiston-Lavier, Laurent Duret et Gwenaël Piganeau.

Spéciale dédicace à toutes les personnes que je côtoie ou que j'ai côtoyées pendant ce doctorat. les copains du master BCD avec qui je suis toujours en contact : Hugo, Florian, Gonché, Abdou, Marianne, Quentin ; les copains Jean-Benoît et Julien (et à toutes les personnes que j'ai pu oublier). Un immense *big up* à tous les BoyZ : Nathan, Lucas, Toscan, Pierre-Yves, Samuel, Théo et affilié(e)s. Au top of the pop, vous avez toujours été là pour moi et m'avez particulièrement soutenu dans cette aventure. J'ai passé des moments inoubliables avec vous, et chacune de nos retrouvailles est un pur moment de joie qui égaye ma semaine.

Je souhaiter dédier cette thèse à tous les membres de ma famille et à tous ceux que j'aime. Merci à mes parents, qui ont toujours été un exemple pour moi de rigueur intellectuelle et de curiosité scientifique et littéraire, et dont l'éducation sans faille m'a donné le goût de tout. Sans eux, je n'aurais jamais imaginé faire un doctorat. Votre curiosité pour mon sujet (quel courage !) et votre soutien ont été précieux. Merci maman, pour ta vie qui est un poème formidable, une ode à la vie et à l'amour. Merci de m'avoir supporté pendant les moments durs, et de m'avoir apporté tant de choses. Merci papa pour tout ton amour, pour nos incroyables discussions et pour m'avoir fait découvrir le monde universitaire. La poussière des craies, l'ambiance des bureaux et des salles de cours, les pique-niques et nos folles courses à dos de chaises à roulettes dans les sombres couloirs des bâtiments de l'Université de Paul-Valéry m'ont fait réaliser que le milieu universitaire est peut-être ce qui se rapprocherait le plus de ma définition de paradis sur Terre. Un immense merci à mes frères et sœurs Vincent, Xavier, Victor, Hadrien, Martí et Anaïs pour leur amour, leur soutien et tous les bons moments qu'on a passés ensemble. Et merci pour votre intérêt dans mes travaux, j'espère que ce doctorat vous aidera à mieux les comprendre (car je pense que je vous ai pas mal tanné avec ça) ! Plein de bisous à mes oncles, tantes et cousins ! Un gigantesque merci à ma compagne Laura. Nous avons passé ensemble cette épreuve du feu qu'est le doctorat, et mon amour pour toi on est sorti encore plus grand. Merci pour tous ces moments indescriptibles où je me sentais moi-même, pour nos fous-rires et pour nos belles et profondes discussions. Quand je suis avec toi, je suis en paix et je me sens revivre. Les mots ne pourraient pas décrire tout ce que tu me donnes et le bonheur que tu m'apportes. De gros bisous à Georges, Nathan et Zahra, à qui je pense très fort ! Pour finir, je tiens à dédier particulièrement ce doctorat à mes grands-parents Jeanne Saboret et Gilbert Bourret, qui ont grandement participé à la formation de l'être que je suis aujourd'hui. Grand-mère, pour m'avoir suivi tout au long de mes 28 premières années, m'éduquant avec une douce austérité envers un monde que j'avais parfois du mal à saisir. J'ai toujours le goût en bouche du veau, du gratin de chou-fleur, des gnocchis, des gâteaux de chez Houdeman, mais surtout de nos discussions qui furent parmi les plus belles de ma vie. Tu auras toujours été le meilleur exemple de quelqu'un de foncièrement bon, ton cheminement de vie et ton incroyable sacerdoce à l'appui. Papy, merci pour ton humour pince-sans-rire et pour tes histoires incroyables. je me souviendrai toujours de ton leitmotiv ô combien motivant « et quand est-ce que tu découvriras une nouvelle petite bête et que tu lui donneras ton nom ? » . Bientôt papy, bientôt, c'est promis. Vous me manquez terriblement, mais votre souvenir éclairerait le plus sombre des endroits.

Abstract

During the cellular translation process, the ribosomal machinery synthesizes a protein through the successive reading of codons along the messenger RNA. With each codon read, the ribosomes call upon the transfer RNAs, which are loaded with an amino acid (the basic unit of proteins). The complementarity between the codon on the mRNA and the anticodon on the tRNA is evaluated, eventually leading to the polymerization of the amino acid onto the nascent protein. There are 64 codons classically associated with 20 amino acids. Several codons, qualified as synonyms, can thus be associated with the same amino acid. Codon use bias (CUB) refers to the differential use of synonymous codons at the gene, genomic region or genome scale. CUB can be associated with mutational processes, at the origin of local peculiarities of nucleotide composition, but also with selection processes improving the protein synthesis dynamics. The influence of these two processes on CUB has been demonstrated in prokaryotes and in some eukaryotes. However, there is no strong evidence of selection acting at the Vertebrate gene level, and more specifically in mammals. Do we consider the CUB in these species from a correct angle ? Do we have the necessary tools to draw such conclusions ? To answer these questions, we propose a mathematical, computational and analytical approach to CUB through the analysis of vertebrate paralogs and human viruses. We have designed a new CUB index called COUSIN (COdon Usage Similarity INdex), which quantifies the distance between the CUB of a sequence and that of a reference, and which stands out from other existing indices by its clarity in the interpretation of results. This index is implemented within an eponymous tool (http ://cousin.ird.fr). In a second step, we performed a study of the evolutionary history and CUB of the Vertebrates paralogous genes Polypyrimidine Tract-Binding Protein (PTBP) whose tissue-specific expression could be associated with differences in their CUB. We show that PTBP1 paralogs appear to be mutagenically biased towards GC enrichment, while the CUB of PTBP2 paralogs may reflect a translational selection towards the use of rare codons in the genome. We interpret that the evolution of the CUB of PTBPs is compatible with a scenario of paralog sub-functionalization by differential expression during vertebrate development. Finally, we studied CUB in human viruses through human polyomaviruses (PyVs). Due to their obligate parasitism on the cellular protein synthesis machinery, the CUB of viruses could impact the clinical presentation of the infection. Our choice of human PyVs comes precisely from their genotypic diversity as well as their multiplicity of clinical manifestations. Infections with human PyVs are highly prevalent and asymptomatic but, in a context of immunosuppression, they can cause heavy and sometimes fatal tissue symptoms. Polyomavirus BK (BKPyV) is known to cause nephropathy in kidney transplant recipients. To prepare the analysis of longitudinal viremia, viruria and genetic data on kidney transplant patients, we have built two pipelines performing an analysis of the PyV genomes, and in particular of their genotype. In order to better understand the evolutionary dynamics of BKPyV in kidney disease, we have analyzed the evolution and CUB of PyVs in the context of the host-parasite relationship. The results proposed in this thesis enrich the basis for the study of CUB in vertebrates, and foster the debate on the pertinence of tissue-specific analysis through differential gene expression and virus tropism.

Keywords : Evolution ; codon usage bias ; mutational bias ; translation selection ; paralogy ; polyomaviruses

Résumé

Au cours du processus cellulaire de traduction, la machinerie ribosomale synthétise une protéine au travers de la lecture successive des codons le long de l'ARN messager. À chaque codon lu, les ribosomes font appel aux ARN de transfert, chargés d'un acide aminé (l'unité de base des protéines). La complémentarité entre le codon de l'ARNm et l'anticodon de l'ARNt est évaluée, conduisant éventuellement à la polymérisation de l'acide aminé sur la protéine naissante. Il existe 64 codons associés classiquement à 20 acides aminés. Plusieurs codons, qualifiés de synonymes, peuvent donc être associés à un même acide aminé. Le biais d'usage des codons (CUB) désigne l'usage différentiel des codons synonymes à l'échelle d'un gène, d'une région génomique ou d'un génome. Le CUB peut être associé à des processus mutationnels, à l'origine de particularités locales de composition nucléotidique, mais aussi à des processus de sélection pour améliorer la dynamique de synthèse de protéines. L'influence de ces deux processus sur le CUB a été démontrée chez les procaryotes et chez certains eucaryotes. Cependant, il n'existe pas d'évidence forte d'une sélection agissant à l'échelle des gènes des Vertébrés, et plus précisément des mammifères. Considérons-nous donc le CUB dans ces espèces sous un angle correct ? Possédons-nous les outils nécessaires pour tirer de telles conclusions ? Pour répondre à ces questions, nous proposons ici une approche mathématique, informatique et analytique du CUB par le biais de l'analyse de paralogues et de virus humains. Nous avons conçu un nouvel indice de mesure du CUB appelé COUSIN (COdon Usage Similarity INdex), qui quantifie la distance entre le CUB d'une séquence et celui d'une référence, et qui se démarque des autres indices existants par sa clarté dans l'interprétation des résultats. Cet indice est implémenté au sein d'un un outil éponyme (http ://cousin.ird.fr). Dans un deuxième temps, nous avons effectué une étude de l'histoire évolutive et du CUB des gènes paralogues de Vertébrés Polypyrimidine Tract-Binding Protein (PTBP) dont l'expression tissu-spécifique pourrait être associée aux différences dans leur CUB. Nous montrons que les paralogues PTBP1 semblent soumis à un biais mutationnel vers un enrichissement en GC, alors que le CUB des PTBP2 pourrait refléter une sélection traductionnelle vers l'utilisation de codons rares dans le génome. Nous interprétons que l'évolution du CUB des PTBPs est compatible avec un scénario de sous-fonctionnalisation des paralogues par expression différentielle pendant le développement des Vertébrés. Finalement, nous avons étudié le CUB chez des virus humains au travers des polyomavirus humains (PyVs). Du fait de leur mode de vie obligatoirement parasite de la machinerie de traduction, le CUB des virus pourrait impacter la présentation clinique de l'infection. Notre choix des PyVs humains vient précisément de leur diversité génotypique ainsi que de leur multiplicité de manifestations cliniques. Les infections par PyVs humains sont fortement prévalentes et asymptomatiques mais, dans un contexte d'immunosuppression, elles peuvent provoquer des symptômes tissulaires importants et parfois mortels. Le Polyomavirus BK (BKPyV) est notamment connu pour provoquer des néphropathies chez des patients receveurs d'une greffe de reins. Pour préparer l'analyse de données longitudinales de virémie, de virurie et génétiques sur des patients receveurs d'une greffe de reins, nous avons construit deux pipelines permettant une analyse du génome des PyVs et de leur génotype. Afin de mieux comprendre la dynamique évolutive de BKPyV dans le cadre des néphropathies, nous avons analysé l'évolution et le CUB des PyVs dans le contexte de la relation hôte-parasite. Les résultats proposés au sein de cette thèse enrichissent les bases pour l'étude du CUB chez les Vertébrés, et alimentent le débat sur les approches tissu-spécifiques au travers de l'expression différentielle des gènes et du tropisme des virus.

Mots-clés : Évolution ; biais d'usage des codons ; biais mutationnel ; sélection traductionnelle ; paralogie ; polyomavirus