HAL
open science

Évolution des préférences d'usage de codons et manipulation de la fidelité de la traduction

Fanni Borveto

▶ **To cite this version:**

Fanni Borveto. Évolution des préférences d'usage de codons et manipulation de la fidelité de la traduction. Sciences agricoles. Université Montpellier, 2021. Français. NNT : 2021MONTT082 . tel-03615085

HAL Id: tel-03615085

https://theses.hal.science/tel-03615085v1

Submitted on 21 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

En Biologie de l'Évolution

École doctorale CBS2

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

Évolution des préférences d'usage de codons et manipulation de la fidelité de la traduction

Présentée par **Fanni BORVETŐ**
Le 10 Décembre 2021

Sous la direction de Ignacio González BRAVO

Devant le jury composé de

Ingrid LAFONTAINE, Professeure, SU, Paris	**Rapporteure**
Daniel M. WEINREICH, Professor, Brown University, Providence	**Rapporteur**
Céline SCORNAVACCA, Directrice de Recherce, CNRS, Montpellier	**Examinatrice**
Lucie ÉTIENNE, Chargée de recherche, CNRS, Lyon	**Examinatrice**
Luis-Miguel CHEVIN , Directeur de Recherche, CNRS, Montpellier	**Examinateur**
Ignacio González BRAVO, Directeur de Recherche, CNRS, Montpellier	**Directeur de thèse**

UNIVERSITÉ DE MONTPELLIER

Evolution of Codon Usage Preferences and Manipulation of Translation Accuracy

Presented by **Fanni BORVETŐ**

10 Decembre 2021

Supervised by Ignacio González BRAVO

Members of the Jury

Ingrid LAFONTAINE, Professeure, SU, Paris	**Reviewer**
Daniel M. WEINREICH, Professor, Brown University, Providence	**Reviewer**
Céline SCORNAVACCA, Directrice de Recherce, CNRS, Montpellier	**Jury**
Lucie ÉTIENNE, Chargée de recherche, CNRS, Lyon	**Jury**
Luis-Miguel CHEVIN , Directeur de Recherche, CNRS, Montpellier	**Jury**
Ignacio González BRAVO, Directeur de Recherche, CNRS, Montpellier	**Thesis supervisor**

UNIVERSITÉ DE MONTPELLIER

Abstract

Eukaryotic cells contain a complex cellular machinery, that regulates and carries out gene expression. The standard genetic code that is the basis of this protein production line is redundant, meaning that 64 codons encode for 20 amino acids. This redundancy gives rise to synonymous codons, that encode for the same amino acid. Synonymous codons are not used at random, genes, tissues and organisms tend to have divergent Codon Usage Preferences (CUPrefs). The role of CUPrefs and the forces that shape them are not yet clear, although it is certain that they hold an important regulatory position in gene expression. If a gene's CUPrefs match the cellular tRNA pool, translation will be fast and efficient, while under- or overmatching CUPrefs may cause either slow and inaccurate translation or competition among genes for resources. Viruses are dependent of the host cell's resources to express their genes, therefore the study of their CUPrefs is primordial to understand their functioning and interactions with the host. In this work, we attempt to enlarge our understanding of the importance of CUPrefs by analyzing the causes and consequences of CUPrefs in eukaryotes and viruses, and in a long-term evolution experiment.

First, we analyzed eleven recombinant Papillomaviruses (PV) that infect exclusively Cetaceans, along with other PVs that infect the same host order: the Cetartiodactyles. We found that recombinant PVs, are not different from non-recombinants in terms of CUPrefs. Instead CUPrefs are associated to gene type, with a link to gene function, and expression pattern. They do not match host CUPrefs, hinting to an immune evasion strategy by keeping low viral gene expression due to the undermatch. Next, we looked at the evolution of CUPrefs in the three paralogs in vertebrates encoding for the Polypyrimidin tract binding protein (PTBP). The PTBP paralogs show distinct CUPrefs, with a GC enrichment linked to local mutational forces in PTBP1 in mammals. We propose that the divergent nucleotide composition in PTBPs is a result of evolution by sub-functionalisation upon gene duplication, and that it's linked to gene expression patterns in different tissues. In an experimental evolution setup we introduced synonymous genes (that only differ in CUPrefs) under strong selection for expression into HEK293 cells, and let them evolve under three conditions for a hundred generations. When the heterologous genes under are directly under selection, cells overcome CUPrefs mismatch, and in spite of the differences, converge to a similar expression pattern. In contrast, when the modified genes are subject of genetic hitchhiking, regulatory mechanisms lead to different expression profiles to limit metabolic cost.

Overall we show that the CUPrefs play a role in regulating gene expression in terms of its differed time or place. Further, we suggest that Eukaryote cells can adjust rapidly by complex regulatory mechanisms to overcome the disadvantages of heterologous CUPrefs if they are needed for survival, or down-regulate them if their expression is costly.

Résumé

Les cellules eucaryotes contiennent une machinerie cellulaire complexe, qui contrôle l'expression des gènes. Le code génétique qui est à la base de cette ligne de production est dégénéré, ce qui signifie que 64 codons (trois bases consécutives) codent pour 24 acides aminés. Cela donne lieu à des codons synonymes, qui codent pour le même acide aminé. Les codons synonymes ne sont pas utilisés au hasard, les gènes, les tissus et les organismes ont tendance à avoir des préférences d'utilisation des codons (CUPrefs) divergentes. Le rôle des CUPrefs et les forces qui les façonnent ne sont pas encore clairs, mais il est certain qu'ils occupent une position importante dans l'expression des gènes. Si les CUPrefs d'un gène correspondent au pool d'ARNt, la traduction sera rapide et efficace, tandis qu'une correspondance insuffisante ou excessive des CUPrefs peut entraîner une traduction lente et imprécise ou une compétition entre les gènes pour des ressources telles que les ARNt et les ribosomes. Les virus sont dépendants des ressources de la cellule hôte pour exprimer leurs gènes, l'étude de leurs CUPrefs est donc primordiale pour comprendre leur fonctionnement et leurs interactions avec l'hôte. Dans ce travail, nous tentons d'élargir notre compréhension de l'importance des CUPrefs en analysant les causes et les conséquences des CUPrefs chez les eucaryotes et les virus, et dans une expérience d'évolution à long terme.

Pour commencer, nous avons analysé 11 Papillomavirus (PV) recombinants qui infectent exclusivement des Cétacés, ainsi que d'autres PV qui infectent le même ordre d'hôtes : les Cetartiodactyles. Nous avons constaté que les PV recombinants ne sont pas différents des non-recombinants en termes de CUPrefs. Au contraire, les CUPrefs sont associés au type de gène. Elles ne correspondent pas non plus aux CUPrefs de l'hôte, ce qui laisse supposer une stratégie d'évasion immunitaire consistant à maintenir une faible expression des protéines virales du au décalage entre les CUPrefs.

Ensuite, nous avons examiné l'évolution des CUPrefs dans le Polypyrimidin tract binding protéin (PTBP) et ses trois paralogues chez les vertébrés. Ces paralogues présentent des CUPrefs distincts, avec un enrichissement en GC lié à des forces mutationnelles locales dans PTBP1 chez les mammifères. Nous proposons que la composition nucléotidique divergente des PTBP est le résultat d'une évolution par sous-fonctionnalisation lors de la duplication des gènes, et qu'elle est liée aux modèles d'expression des gènes dans différents tissus.

Dans une manip d'évolution expérimentale, nous avons introduit des gènes synonymes (qui ne diffèrent que par leurs CUPrefs) dans des cellules HEK293, et nous les avons laissé évoluer sous trois types de traitement pendant une centaine de générations. Nous avons constaté que lorsque les gènes hétérologues sont directement soumis à la sélection, les cellules surmontent le décalage des CUPrefs et, malgré les différences, convergent vers un modèle d'expression similaire. En revanche,

lorsque les gènes modifiés font l'objet du hitchiking génétique, les mécanismes de régulation conduisent à des profils d'expression différents afin de limiter le coût métabolique.

Dans l'ensemble, nous avons constaté que les CUPrefs jouent un rôle dans la régulation de l'expression des gènes en fonction du moment ou du lieu de leurs expression, comme on l'observe à la fois chez les PVs de et chez les vertébrés. Pendant ce temps, les cellules eucaryotes peuvent s'adapter rapidement par des mécanismes de régulation complexes pour surmonter les désavantages des CUPrefs hétérologues s'ils sont nécessaires à la survie, ou les inhiber si leur expression est coûteuse.

Résumé long

Les cellules eucaryotes contiennent une machinerie cellulaire complexe, qui contrôle l'expression des gènes. Le code génétique qui est à la base de cette ligne de production est dégénéré, ce qui signifie que 64 codons (trois bases consécutives) codent pour 24 acides aminés. Cela donne lieu à des codons synonymes, qui codent pour le même acide aminé. Les codons synonymes ne sont pas utilisés au hasard, les gènes, les tissus et les organismes ont tendance à avoir des préférences d'utilisation des codons (CUPrefs) divergentes. Le rôle des CUPrefs et les forces qui les façonnent ne sont pas encore clairs, mais il est certain qu'ils occupent une position importante dans l'expression des gènes. Si les CUPrefs d'un gène correspondent au pool d'ARNt, la traduction sera rapide et efficace, tandis qu'une correspondance insuffisante ou excessive des CUPrefs peut entraîner une traduction lente et imprécise ou une compétition entre les gènes pour des ressources telles que les ARNt et les ribosomes. Les virus sont dépendants des ressources de la cellule hôte pour exprimer leurs gènes, l'étude de leurs CUPrefs est donc primordiale pour comprendre leur fonctionnement et leurs interactions avec l'hôte. Pour cela, nous devons d'abord comprendre comment une cellule hôte exprime des gènes hétérologues qui peuvent ou non correspondre aux CUPrefs de la cellule. Cette question a commencé à être explorée chez les bactéries, mais beaucoup moins chez les eucaryotes, et notamment chez les mammifères. Globalement, les cellules eucaryotes ont une chaîne de production de protéines bien plus complexe que les procaryotes. Outre le cloisonnement des structures et des fonctions cellulaires, on retrouve des processus de régulation, des mécanismes de relecture et de correction qui assurent le potentiel de modulation de l'expression des gènes.

Dans ce travail, nous tentons d'élargir notre compréhension de l'importance des CUPrefs en analysant les causes et les conséquences des CUPrefs chez les eucaryotes et les virus, et dans une expérience d'évolution à long terme.

Pour commencer, nous avons analysé 11 Papillomavirus (PV) recombinants qui infectent exclusivement des Cétacés, ainsi que d'autres PV qui infectent le même ordre d'hôtes : les Cetartiodactyles. La particularité de ces PVs recombinants, est que leur région précoce appartient au groupe des PVs Alpha-Omikron alors que la région du gène tardif appartient à celle du groupe Beta-Xi. Nous avons collecté tous les génomes de PV infectant des Cetartiodactyles, avec leurs métadonnées correspondantes sur l'espèce hôte, la localisation anatomique et la présentation clinique. Notre ensemble de données regroupe 58 PV de trois groupes différents, qui infectent 20 espèces hôtes différentes.

Après l’analyse des motifs régulateurs, des CUPrefs et de la reconstruction phylogénétique de ces PVs, nous proposons qu'un seul événement de recombinaison se trouve à l'origine de ce groupe recombinant, ce qui leur a permis d'évoluer par la suite vers un ensemble de caractéristiques uniques. Les résultats suggèrent que les motifs régulateurs identifiés ne sont pas une combinaison de ceux des lignées parentales mais sont spécifiques au groupe recombinant. La PV recombinante la plus basale ne présente pas ces motifs conservés, ce qui suggère en outre qu'ils sont apparus comme une adaptation à des nouvelles conditions. En outre, la distribution des motifs régulateurs est bien corrélée non seulement avec la taxonomie virale mais aussi avec celle de l'espèce hôte. Cela pourrait être interprété comme une adaptation réglementaire à l'hôte.
En analysant les CUPrefs des PVs de Cetartiodactyla, nous n'avons pas observé de différences significatives entre les PVs recombinants et non-recombinants, nous avons plutôt constaté que les CUPrefs dépendent du type de gène et donc du moment d'expression au cours de l'histoire naturelle de l'infection. En effet, les gènes Early et Late ont des CUPrefs facilement différenciables, ce qui pourrait correspondre aux différents stades de vie d'un kératinocyte où ils sont exprimés. Nous avons également comparé les CUPrefs des PV à ceux de leurs hôtes à l'aide de l'outil COUSIN. Nous avons observé que, comme c'est le cas pour d'autres PV, les virus étudiés ne correspondent pas aux CUPrefs de leurs hôtes respectifs, et présentent au contraire des CUPrefs opposés compatibles avec une stratégie d'évasion du système immunitaire. Cette hypothèse est cohérente avec les observations faites chez d'autres virus, où les CUPrefs de leurs gènes semblent être une adaptation à l'environnement de l'hôte en évitant la réponse immunitaire.

Ensuite, nous avons examiné l'évolution des CUPrefs dans le Polypyrimidin tract binding protéin (PTBP) et ses trois paralogues chez les vertébrés. Nous proposons qu'au fil du temps, les gènes paralogues peuvent évoluer pour avoir des CUPrefs divergents, qui permettent leur expression différentielle dans l'espace et le temps. Nous utilisons l'exemple des PTBP (Polypyrimidin tract binding proteins), codées par un certain nombre de gènes présents chez tous les vertébrés. Ces gènes sont présents sous forme de trois paralogues principaux, PTBP1, PTBP2 et PTBP3. Robinson et ses collaborateurs ont montré que chez l'homme, ces paralogues ont des profils d'expression différents liés à leurs différentes CUPrefs ((Robinson, Jackson, & Smith, 2008)). Cette évolution différentielle de la synchronisation de l'expression des gènes pourrait être interprétée comme le résultat de pressions sélectives libérées sur le gène dupliqué, comme l'original peut toujours conserver sa fonction, tandis que le duplicata peut explorer de nouvelles fonctions ou de nouveaux profils d'expression.

Nous avons étudié les paralogues de PTBP de 47 vertébrés mammifères et 27 vertébrés non-mammifères, ainsi que trois espèces de protostomes en tant qu'outgroupes. Quinze de ces espèces de vertébrés ont été étudiées plus en profondeur car nous avons pu récupérer des génomes complets bien annotés. Nous avons utilisé des méthodes de regroupement, de reconstruction phylogénétique, de reconstruction de l'état ancestral, et nous avons examiné les forces qui ont pu façonner les CUPrefs de ces gènes.

Nous avons observé que les PTBP1s sont les paralogues les plus riches en GC, tandis que les PTBP3s sont les plus riches en AT. De plus, il existe une différence significative dans le contenu GC entre les PTBP1s des vertébrés mammifères et non-mammifères, les espèces mammifères étant enrichies en GC. Ceci est confirmé par clustering hiérarchique. En analysant les CUPrefs complets, nous retrouvons trois groupes : PTBP1 chez les mammifères, PTBP1 chez les non-mammifères, et tous les PTBP2 & PTBP3 se regroupant ensemble.

Dans les 15 espèces bien annotées, nous avons inspecté le contexte génomique afin d'évaluer si la richesse en GC observée chez les mammifères est le résultat de forces mutationnelles locales. Nous avons donc comparé le contenu GC3 des paralogues à leurs régions flanquantes et à leurs introns. Nous avons constaté que les variations du contexte génomique local expliquent presque complètement les variations du contenu GC3 de PTBP1 (régression séquentielle des moindres carrés, R^2=0.97), relativement bien dans le cas de PTBP2 (R^2=0.46), et moins de la moitié dans le cas de PTBP3 (R^2=0.16). Nous interprétons, que dans le cas des mammifères, il y a un enrichissement GC global qui est clairement visible dans PTBP1, mais qui n'explique pas les CUPrefs de PTBP2 et PTP3. En fait, les valeurs COUSIN des PTBP des mammifères montrent que les PTBP1 dépassent les CUPrefs des organismes, tandis que les PTBP2 et PTBP3 les sous-estiment, ce qui signifie qu'ils ont une fréquence accrue de codons rares. A leur tour, les PTBP non mammaliennes présentent des CUPrefs correspondant à leurs organismes.

La reconstruction phylogénétique regroupe les séquences d'abord par paralogie, puis par espèce, ce qui laisse supposer que deux événements de duplication ont eu lieu au moment de l'émergence des vertébrés. La reconstruction ancestrale et l'analyse des séquences montrent que les mammifères ont accumulé un grand nombre de mutations synonymes et non-synonymes qui enrichissent les séquences en GC, par rapport aux séquences non-mammifères.

Nous avons exploré la nature de l'évolution des gènes paralogues et de leurs CUPrefs, et montré que les PTBP présentent une composition nucléotidique et des CUPrefs divergents sur le clade des vertébrés. Nous proposons que ce phénomène soit compatible avec la théorie de l'évolution génotypique par sous-fonctionnalisation lors de la duplication des gènes.

Finalement, dans une manip d'évolution expérimentale, nous avons lancé une expérience de sélection à long terme, en utilisant un ensemble de versions de gènes synonymes pour mieux comprendre les effets immédiats et à long terme des CUPrefs sur l'expression des gènes chez les eucaryotes, et comment chaque étape du processus est liée. Comme il s'agit d'une expérience à long terme, nous avons également cherché à savoir comment les cellules compenseraient éventuellement une correspondance non optimale d'un gène dont elles ont besoin pour survivre.

Nous avons cloné cinq versions différentes du gène *shble* connecté par un peptide P2A à un gène de protéine fluorescente verte améliorée (*egfp*), dans un plasmide et les avons transfectées dans des cellules HEK293, les rendant résistantes aux antibiotiques. Elles ont été soumises à trois traitements de sélection différents : la Bléomycine, un antibiotique auquel elles peuvent résister en exprimant le gène synonyme *shble* correspondant ; la Néomycine, un antibiotique auquel les cellules peuvent résister en exprimant le gène *neo_tp* non modifié, présent dans les plasmide clonées; et en l'absence d'antibiotiques dans les milieux. Le complexe *shble-egfp* est sous le contrôle d'un puissant promoteur du *cytomégalovirus* (CMV), assurant un niveau de transcription élevé au lancement de l'expérience. L'expression des gènes a été contrôlée en suivant les niveaux d'ADN, d'ARNm et de protéines, tandis que le phénotype cellulaire a été contrôlé en quantifiant l'intensité de la fluorescence et la capacité des cellules à se développer en présence d'antibiotiques.

Nous avons essayé d’explorer en profondeur les effets de la CUPref de gènes hétérologues dans les cellules eucaryotes, et comment les cellules peuvent compenser la charge différentielle imposée par l'expression de ces différents gènes synonymes sur 100 générations. Malgré les défis de l'évolution expérimentale à long terme et la complexité de notre ensemble de données multi-niveaux, nous avons pu identifier un certain impact des CUPrefs sur les cellules. Nous montrons que si l'expression des gènes modifiés est directement soumise à la sélection, les cellules surmontent sans coût notable la discordance des CUPrefs et, malgré les différences, convergent vers des modèles d'expression similaires. En revanche, lorsque les gènes modifiés sont soumis à un auto-stop génétique, des mécanismes de régulation potentiels créent des profils d'expression différents pour limiter le coût métabolique, jusqu'à inhiber complètement la traduction et probablement la transcription du gène en question.

Bien qu'il y ait un débat sur les forces qui façonnent les CUPrefs chez les vertébrés, nous avons découvert que, selon la fonction des gènes et le modèle d'expression dans le temps et l'espace, plusieurs facteurs façonnent les CUPrefs chez les eucaryotes supérieurs et que cela ne peut pas être expliqué par un simple biais mutationnel ou une simple sélection translationnelle. Le biais de mutation locale, la conversion génétique basée sur le GC et la sélection translationnelle agissent

tous les -deux dans le cas de gènes paralogues- sur plusieurs millions d'années, conférant aux paralogues des CUPrefs distincts. Cependant, l'évolution expérimentale montre qu'à une échelle de temps plus petite, ces forces sont négligeables, mais que les processus de régulation épigénétique sont rapides pour ajuster les modèles d'expression, si nécessaire, malgré et indépendamment des CUPrefs. Lorsque la mutation et la sélection translationnelle ont le temps et le pouvoir d'agir sur une population suffisamment importante, tant chez les virus que chez les vertébrés, elles peuvent façonner les CUPrefs d'un gène pour l'adapter à son profil d'expression et à sa fonction. Par exemple, dans les PTBP, il existe des signes de spécificité tissulaire, et les CUPrefs de chaque paralogue semblent suivre les CUPrefs de son environnement (voir chapitre 3). Dans les papillomavirus, nous observons des CUPrefs divergentes entre les gènes qui sont exprimés au stade précoce et au stade tardif de l'infection (voir chapitre 2). Dans les mêmes PVs, nous avons observé une corrélation entre la présence de motifs régulateurs conservés et l'espèce hôte. Cela pourrait indiquer une coévolution entre l'hôte et l'initiation de la transcription et de la traduction par le pathogène, car les eucaryotes semblent intervenir efficacement dans l'expression de gènes hétérologues. En revanche, si l'expression (ou la non-expression) d'un gène hétérologue ne constitue pas une menace immédiate, ou est trop coûteuse sur le plan métabolique, avec un promoteur fort et les bons motifs de régulation, il peut quand même être exprimé (voir chapitre 4).
Les CUPrefs des virus infectant les vertébrés vont généralement à l'encontre du biais des codons de l'hôte, et il a été proposé que c'est en partie pour éviter le système immunitaire de l'hôte. Dans notre expérience de sélection, les cellules ne disposaient pas d'un système immunitaire adaptatif, mais dans le cas de Shble1, les cellules ont partiellement réduit au silence la transcription du complexe shble-egfp surajouté. Nous proposons que, même en l'absence d'une réponse immunitaire, les cellules peuvent épigénétiquement réguler à la baisse les gènes viraux s'ils hébergent un coût métabolique élevé immédiat. Pendant ce temps, les gènes viraux dont les CUPrefs ne correspondent pas peuvent être exprimés en arrière-plan sans alarmer davantage le système immunitaire ou les mécanismes de régulation de la cellule hôte.

Dans l'ensemble, nous avons constaté que les CUPrefs jouent un rôle dans la régulation de l'expression des gènes en fonction du moment ou du lieu de leurs expression, comme on l'observe à la fois chez les PVs de et chez les vertébrés. Pendant ce temps, les cellules eucaryotes peuvent s'adapter rapidement par des mécanismes de régulation complexes pour surmonter les désavantages des CUPrefs hétérologues s'ils sont nécessaires à la survie, ou les inhiber si leur expression est coûteuse.

Acknowledgments

First of all, I'd like to thank Ingrid Lafontaine and Daniel Weinreich to have accepted to read and correct my manuscript. I would also like to thank Céline Scornavacca, Lucie Étienne and Luis-Miguel Chevin to have accepted to be the judges of my work.
I'd like to thank my follow up committee : Camille Martinand-Mari, Frédéric Delsuc, Alison Duncan and Olivier Duron, to have been so kind and encouraging both scientifically and in their interactions. Your good advice and our discussions helped us to see clear in the total mess our data were sometimes.

And of course a huge thanks to you Nacho, who, for whatever mysterious reason accepted me as an intern and then as a PhD student. I have learned so many MANY things thanks to you, about science, philosophy and a thousand random facts about everything that exists, existed or may exist in a parallel universe. Oh, and a few things about codons and Papillomaviruses of course. I had many doubts about being able to survive this, but you have always managed to motivate me, and I'm truly thankful for your trust in me.
I am probably the luckiest PhD student, as the many people who turned up in our team were all not just excellent scientists, but also the most awesome! I'd like to thank you the entire Virostyle team for all their help and corridor discussions! Thank you Jérome to have helped get an internship, without you, I really wouldn't be here! And for your patience when I was asking a thousand questions about R and everything, because I was always lost. I'd like to thank Anouk, who supervised my internship and helped me publish my very first article, but was also like a scientific fairy godmother, with infinite many good advice on life in the scientific field. I'd like to thank Antonin for teaching me to be super clean in my labwork, and for his patience and advice while correcting my presentations. Marion and Laura, both of you were watching over me like you were my big sisters, and I cannot thank you enough. I would also like to thank Cécile for her help, she knows the how and where of everything in the lab, and she always helped us to find the things we needed.
And there was this other person but I'm not sure of her name, I rarely saw her…oh no wait, she's my new neighbor! Fiona! You would deserve an entire book of thank yous!

First, you single-handedly maintained the trillion cells; you were ready to brake into the lab in the middle of lockdown to keep them alive, so needless to say, you 100% deserve the title of co-PhD! And most importantly, without you and your friendship I would have abandoned all this after a year or so. You can add to your CV the title “PhD student mental health manager”. You always had a carambar, or a “scrounch” to feed a starving Fanni, or an escape-office-game to welcome me back from vacation and the list goes on and on, I tell you the rest tomorrow! So just…. Thank you! Kösziiii! Merciiiiiii avec pleins de iiiiiiiiiii :D

Yann, you will have a whole paragraph too! I think you were more stressed than me about this, and despite that you helped me to get through it! You were there to reassure me when I was crying because I didn’t believe in myself, and you could explain me, why I am wrong to believe that. Which, then always made sense and helped me get through it. So thank you! You also kept cooking all the best home made meals while I was working and I really hope you will keep doing this. We can now go to restaurants and watch movies, I finished it! Je t’aime beaucoup beaucoup Yann!

And last but not least I am infinitely grateful to my parents, who were far far away, but this didn’t stop them from being the most supportive people in all this. From the very beginning when I announced at around age four, that I will be a biologist, you always motivated me and gave your best so that I can achieve my goals. You did everything so that I could go to the University in Montpellier, even though it’s super far from Budapest. I always look forward to our skype sessions, and even more to go home and see you in person! You are the best parents that ever existed on earth and once again I could write an entire novel of thank yous, but I will rather call you in a few minutes.
I hope you are not too angry that I left the cat at home, I’ll will try to get you that seaside house to make up for it!

Table of Contents

Acknowledgments....15
Glossary....20
The basics of Protein synthesis....23
Papillomaviruses infecting cetaceans exhibit signs of genome adaptation following a recombination event....35
Subfunctionalisation of Paralogous Genes and Evolution of Differential Codon Usage Preferences : The Showcase Of Polypyrimidine Tract Binding Proteins....53
The effects of Codon Usage Preferences on Heterologous gene expression in a long-term selection experiment....85
- Introduction....85
- Materials and methods....87
  - Experimental evolution setup....87
  - Transfection, maintenance....91
  - Sampling....94
  - Sample treating....94
  - Day2 experiment....95
- Data preparation and analysis....97
  - Fluorescence Intensity Data....97
  - qPCR, and rt-qPCR Data....103
  - RNAseq Data....104
  - Proteomics Data....105
  - Real-time cell growth measure Data....106
- Results....107
  - Results for Fluorescence Data....108
  - Analyses for mRNA levels through time and selection using rt-qPCR....123
  - Analyses for mRNA levels through time and selection using RNASeq....124
  - Analyses for Protein levels through time and selection using mass spectrometry....129
  - Correlations between molecular steps from genotype to phenotypes....134
  - Real-time cell growth measure....139
- Discussion....140
  - Differences between replicates can be explained by strong bottleneck....141
  - Missing proteins, the limits of Mass Spectrometry....141
  - Transfected cells reach and maintain high resistance to Bleomycin, without apparent fitness cost....143
  - Transcription of the spliced forms of Shble4 and Shble6 is low....143
  - Differences in CUPrefs do not explain variation in transcription and translation levels under Neomycin....144
  - Heavy selective pressure leads to convergent phenotypes under the Bleomycin treatment…146
  - Low correlation between genome, transcriptome and proteome, hints at both pre- and post-transcriptional regulations under Bleomycin....147
  - Cells carrying Shble1 lose *egfp* expression....147
- Conclusion....148
  - Perspectives....148

General Discussion....153
- Conclusion....155

Loop 1 of APOBEC3C Regulates its Antiviral Activity against HIV-1....159
Bibliography....193

Glossary

A - adenine

aaRS - Aminoacyl-tRNA Synthetases

aa-tRNA- Aminoacyl-tRNA

ANOVA - Analysis of variance

Bleo - Bleomycin

C - cytosine

CAI - Codon Adaptation Index

CMV - Cytomegalovirus

COUSIN - Codon Usage Similarity Index

CUB – Codon Usage Bias

CUPrefs – Codon Usage preferences

DENV – Dengue virus

DNA - Deoxyribonucleic acid

downmed – mean of the lowest 20% fluorescent cells

dsDNA - double strain DNA

FACS - Fluorescence Activated Cell Sorting

FBS - Fetal Bovine Serum

FSC- Forward scatter

G - guanine

lhub – log10 of Huber central estimator

MEM- Minimum essential media

mRNA – messenger RNA

Neo - Neomycin

PAS - Polyadenylation signal

PBS - phosphate buffered saline

pre-mRNA – precursor mRNA

PTBP - polypyrimidine tract binding protein

PV – Papillomavirus

qPCR - quantitative real-time PCR

R1- Replicate 1

R2 – Replicate 2

rIBAQ - relative intensity Based Absolute Quantification

RIN – RNA integrity number

RIPA - Radioimmunoprecipitation assay buffer

RNA – ribonucleic acid

RNA pol – RNA polymerase

rt-qPCR- reverse transcription–qPCR

ORF – Open Reading Frame

SSC – Side scatter

T - thymine

TPM – Transcript per million

tRNA- transfer RNA

U - uracil

upmed - mean of the highest 20% fluorescent cells

URR – upstream regulatory region

UTR – Untranslated region

vif – Viral infectivity factor

woAB – without Antibiotic

WT- Wild type

Chapter one

The basics of Protein synthesis

DNA can be transcribed into mRNAs which on their turn are translated into proteins. It is an easy to understand concept known as the Central Dogma, and we may think of the cell as a well oiled chemical machine that works without fault. But of course as most things, when we look deeper in the mechanism, behind the cogs and gears, the clockwork-like precision falls apart. Errors appear here and there, and even today, despite the many works of great scientists we still do not fully understand all the fine-tuning of protein synthesis. In this work we investigate a small series of tiny cogs in the machine, the codon usage preferences and its effects on the protein synthesis in eukaryotic cells. Before leaping in to the study of codons, it is important to learn, and acknowledge what has been already understood about gene expression and its underlying mechanisms, starting with the very beginning : the DNA.

Although DNA as a substance had already been identified in the late 1860s by Friedrich Miescher(Dahm, 2008), it took almost a century of experiments and the work of several now legendary scientist (Chargaff, Vischer, Doniger, Green, & Misani, 1949; F. Crick & Watson, 1953) to identify its hereditary nature, structure, and to eventually crack the code of the notorious giant molecule. It is now well established that in all cellular organisms the DNA consists of two strands that form a double helix, and that the genetic code is made up of not more than four nucleotides : adenine (A), cytosine (C), guanine (G) and thymine (T). These bases go by purine-pyrimidine pairs: A-T and G-C on the opposing strands, and read by triplets also known as codons on the same strand.

The strands of DNA (one at a time) serve as a template for RNA transcription. The RNA is a complementary copy of the DNA with the main differences being containing uracil (U) instead of thymine (T) and Ribose being the main component in place of Deoxyribose. During transcription, DNA is “read” by RNA polymerases where specific promoter sequences recruit them at the beginning of genes (Smale & Kadonaga, 2003).

In this work we focus on protein synthesis in eukaryotes (Figure 1), which is characterized by a strong compartmentalization, meaning that, contrary to procaryotes, the different steps of gene expression happen in different locations in the cell : transcription occurs in the nucleus of a cell, mRNA will be translated in the cytoplasm by the ribosomal machinery.

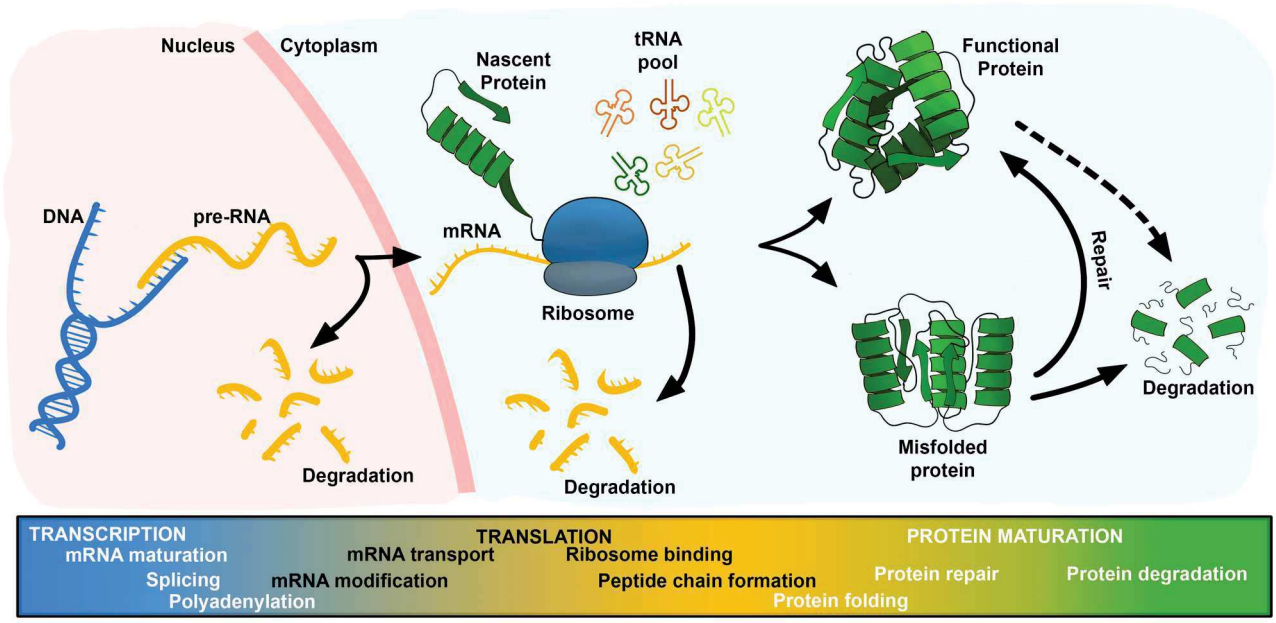


Figure 1: Overview of the gene expression pathway *– simplified representation of the different steps of gene expression. The stages are labeled in the frame below the illustration, from left to right. The first steps happen in the nucleus indicated by the pink background, while translation and protein maturation take place in the cytoplasm represented by a blue background.*

In eukaryotes there are three types of RNA polymerases (RNA pol) while there is only one in prokaryotes. Here we will mostly talk about RNA pol II, which transcribes messenger RNAs (mRNA) that are later translated into proteins. It is worth mentioning, that RNA pol I transcribes ribosomal RNAs, while RNA pol III transcribes mostly tRNAs, and that both RNA molecules have an essential role in translation.

Transcription initiates when upstream an open reading frame a promoter sequence recruits the RNA pol. To engage RNA pol II, housekeeping genes are almost always preceded by long GC rich stretches with CpG islands and the promoter sequence embedded in it (Deaton & Bird, 2011). Genes that are only expressed in specific cell types however, tend to lack this GC rich upstream region. Instead, they have core promoter sequences, that contain conserved elements like the TATA box (Juven-Gershon, Hsu, Theisen, & Kadonaga, 2008). These elements -usually in a cooperative manner- are responsible for initiating transcription. We can also observe enhancer sequences that – contrary to promoter sequences- are not location-specific, but play a very similar role to promoters. They do not interact directly with the polymerase, but are recognized by specific transcription factors that stimulate RNApol II to bind to the promoter(Müller, Gerster, & Schaffner, 1988). Apart from promoters and enhancers, transcription factors are also necessary for efficient transcription. In fact, the presence (or absence) of certain transcription factors explains in part the high difference in the protein repertoire between different cell types (Getzenberg, 1994). The promoter sequences,

transcription factors and the polymerase attached to the DNA create together the pre-initiation complex (PIC), and once everything is in place, the transcription can finally begin. But where does it end? While RNA pol I and III, stop transcription at specific termination sequences, RNA pol II does not seem to have termination sites (or at least nothing has been hitherto identified as such), (Németh et al., 2013; Verosloff et al., 2021). It seems that RNA pol II stops transcription after it has passed a Polyadenylation signal (PAS) that recruits cleaving factors, and that post-transcriptional processing of the precursor mRNA (pre-mRNA) leads also to transcription stop (Eaton & West, 2020). The pre-mRNA is the primary transcript that will later become the mature mRNA. At this point the pre-mRNA is still attached to the polymerase, but cleaving factors will cut the end of the transcript after the highly conserved "AAUAAA" PAS sequence. At the same spot a polyA tail is added by the poly(A) polymerase, this polyA tail will help exporting, maintaining and protecting the mature mRNA (Brown, Valenstein, Yario, Tycowski, & Steitz, 2012; Mitton-Fry, DeGregorio, Wang, Steitz, & Steitz, 2010; Torabi et al., 2021). While the polyA tail is added to the 3' of the mRNA, the 5' end receives a cap while the transcript is growing (Voet, Voet, & Pratt, 2016). This consists of a 7 methylguanosine residue joined to the transcript's initial 5' nucleotide. The 5' cap identifies the translation starting site and has a critical role to ensure mRNA stability and translation efficiency (Jiao et al., 2010; Li & Kiledjian, 2010; Merrick, 2004; Meyer, Temme, & Wahle, 2004).

Eukaryotic pre-mRNAs span multiple unexpressed regions, called introns, that are cut out during mRNA maturation (splicing) (Berget, Moore, & Sharp, 1977; Chow, Gelinas, Broker, & Roberts, 1977; Poverennaya & Roytberg, 2020). Curiously the total length of introns surpasses that of exons (expressed regions) by four to ten folds, creating a mismatch between gene length and protein size in eukaryotic genomes (Hawkin, 1988). Despite being present in high numbers, introns are still full of mystery. In fact, the splicing sites –the boundary locations where introns are cut out- do not seem to be linked to specific sequences. Although there are some recurrent patterns, and available methods for predicting splicing sites are getting more and more precise, there are still many obstacles to overcome before achieving accurate prediction (Ohno, Takeda, & Masuda, 2018). As for the function of the very abundant introns, alternative splicing of the mRNA allows one gene to code for several proteins, and thus act as a rapid and efficient method to increase protein diversity, without introducing mutations to the genome.

After splicing, most mRNAs are ready to be translated into proteins, but in some cases they may undergo additional editing. This may happen simultaneously with other maturation steps and may include chemical edition of bases, insertions, deletions, and even the introduction of new splicing sites (Bentley, 2014; Nishikura, 2010). Once again, the process can be regarded as an ensemble of ways to increase protein diversity without touching the DNA. It was also found that mRNA editing

could regulate protein synthesis, for example methylation of some bases result in the m^6A modification of the mRNA. This m^6A change can affect the binding of proteins on the methylated mRNAs to regulate their translation (Frye, T. Haranda, Behm, & He, 2018; Roundtree, Evans, Pan, & He, 2017). Furthermore, the introduction of new splicing sites in mRNAs present in the nervous system may even modify behavior patterns in mammals and insects (Reenan, 2001).

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	*Phenylalanine*	UCU	*Serine*	UAU	*Tyrosine*	UGU	*Cysteine*	U
	UUC		UCC		UAC		UGC		C
	UUA	*Leucine*	UCA		UAA	*Stop*	UGA	*Stop*	A
	UUG		UCG		UAG	*Stop*	UGG	*Tryptophan*	G
C	CUU		CCU	*Proline*	CAU	*Histidine*	CGU	*Arginine*	U
	CUC		CCC		CAC		CGC		C
	CUU		CCA		CAA	*Glutamine*	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	*Isoleucine*	ACU	*Threonine*	AAU	*Asparagine*	AGU	*Serine*	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	*Lysine*	AGA	*Arginine*	A
	AUG	*Methionine*	ACG		AAG		AGG		G
G	GUU	*Valine*	GCU	*Alanine*	GAU	*Aspartic acid*	GGU	*Glycine*	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	*Glutamic acid*	GGA		A
	GUG		GCG		GAG		GGG		G

Table 1: Standard RNA Codon Table *– The genetic code to translate from the 64 codons of an mRNA to the 20 amino acids. As the code is redundant, multiple codons can code for the same amino acid*

In order to translate from nucleotides to amino acids we need a key to decipher the sequence. This key is the genetic code which converts codons (three consecutive bases) into amino acids (Table 1). The 64 possible codons, encode overall for three STOP codons and 20 amino acids, among which Methionine (ATG) acts as a start codon(Brenner, Stretton, & Kaplan, 1965; F. H. C. Crick, Barnett, Brenner, & Watts-Tobin, 1961). For a while genetic code was thought to be universal, but it has been shown that there are a number of exceptions that use alternative versions, such as in mammalian mitochondria or some ciliated protozoa (Anderson et al., 1981; Jukes & Osawa, 1993).

Mature mRNA, once transported to the cytoplasm, recruits Ribosomes at its 5’ end (Figure 2). This part is not translated in most cases, but serves as a ribosome-binding site, and depending of the length of the untranslated region (UTR), it may contain regulatory sequences that modulate ribosomal affinity and binding. Once a ribosome is recruited, the elongation process may start. Because the mRNA is linear and the interactions with Ribosomes are sequential, several Ribosomes can read the same mRNA simultaneously, increasing the expression rate of the protein. For each codon, the ribosome recruits the matching tRNA with the anticodon and the amino acid it transports.

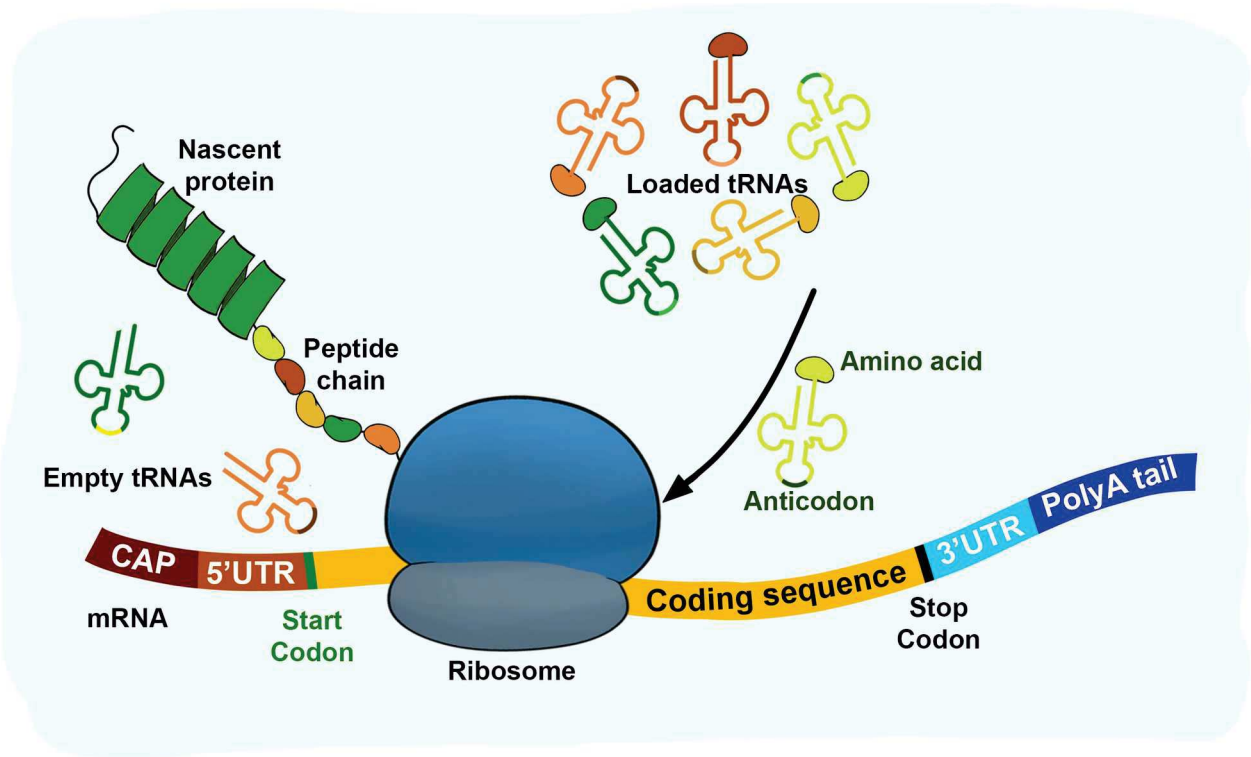


Figure 2: Translation *– simplified representation of the translation process. The Ribosome is connected to the mRNA and is represented in as it is recruiting tRNAs loaded with amino acids (right side) to the corresponding codon in the coding sequence. On the left is the forming peptide chain that is in the process of folding into the nascent protein. The different elements of the mRNA, are also labeled on the sequence (5' Cap, 5' UTR, START codon, Coding sequence, Stop codon, 3' UTR and the polyA tail)*

Transfer-RNAs (tRNA), as their name indicates, are the molecules delivering the amino acids needed for the translation. They are charged beforehand by Aminoacyl-tRNA Synthetases (aaRS), which are accountable for accurate translation. Because Ribosomes don't detect whether the tRNA is loaded with the incorrect amino acid, it is the aaRS that has to bring the right amino acid to the right tRNA, and they are doing a remarkably good job. It was shown that aaRS are special in exhibiting extraordinarily high selectivity (Perona & Hadd, 2012; Tawfik & Gruic-Sovulj, 2020).

Meanwhile tRNAs can recognize more than one codon, to the extent that theoretically 31 different tRNA molecules are enough to translate the 61 codons due to the wobble effect. The wobble hypothesis proposed by Francis Crick, states that the first two codon-anticodon base pairings are strictly determined while the third one allows some limited flexibility (F. H. C. Crick, 1966). Should this be the case, some codon-anticodon pairings could occur that do not necessarily follow Watson-Crick base pair rules (A-U, G-C). These atypical pairings are possible because of modifications in the architecture of the tRNA. In fact the anticodon stem and loop (ASL) often undergoes chemical

changes that either enhance the wobble effect or limit it, such as the introduction of inosine bases, which can pair with all nucleotides except guanine (P. F. Agris, 1991; Paul F. Agris et al., 2018).

As codons are read one after the other in a sequential way triplet by triplet, a peptide chain forms and folds into a protein with the help of chaperons, until the ribosome arrives at a stop codon. As there is no amino acid associated to these codons, the ribosome stops, and release factors disconnect it from the mRNA. Following the detachment of Ribosomes reading it, the mRNA may be degraded or read again by the same or other Ribosomes. Meanwhile tRNAs are loaded by the aaRS with their corresponding amino acids as they are released from the ribosome.

In some cases Ribosomes can pause before reaching a STOP codon. The pause can be induced by different factors, for example : low Aminoacyl-tRNA availability or because of stem-loop formations in the mRNA. In either case, the ribosome may simply be stalled temporally and continue translation later, or stop completely and abandon translation(Buchan & Stansfield, 2007). Therefore, stalling can be a way of regulating expression, but as Ribosomes are limited in a cell it is important to recover them when they don’t follow translation. This is especially needed as the stuck ribosome also blocks all translation complexes upstream on the same mRNA, further limiting resources(Graille & Séraphin, 2012). There are several pathways such as the no-go decay, that degrade faulty mRNA and release of the Ribosome, allowing its recycling(Harigaya & Parker, 2010).

As mentioned above, the 61 codons are translated into 20 amino acids. This means that the genetic code is redundant, and most of the amino acids are encoded by several synonymous codons (Khorana et al., 1966; Nirenberg & J. Heinrich Matthaei, 1961). It has been shown that these synonymous codons are not used at random. In fact, Codon Usage Preferences (CUPrefs) also known as Codon Usage Bias (CUB) – the fact of using one codon over an other synonymous codon- is varying between organisms, tissues, and even along chromosomes and genes (Carbone, Zinovyev, & Képès, 2003; Grantham, Gautier, Gouy, Mercier, & Pavé, 1980; Wada et al., 1990). This means that a cell for example is characterized by an average codon usage, which is determined by the ratio of available tRNAs, however in practice the average CUPrefs are often calculated based on the codons in the genome or transcriptome as it is more accessible than tRNA sequencing. Nonetheless, a gene in the cell with calculated average CUPrefs may possess a matching, undermatching, or overmatching CUPrefs compared to the tRNA pool. A matching CUPrefs means that the frequency of used codons in a gene is close to the proportion of the available tRNAs in its environment, an undermatch on the other hand is when the codons used by the gene in question, are the ones that are rare in the tRNA pool. A gene with an overmatching CUPref uses only the originally most abundant tRNAs. It has been shown that these local variations in CUPrefs between

cells in an organism may modulate protein synthesis, as genes with a preference for rare codons have a slower and lower translation rate, while genes with common codons are the ones that are most expressed (Ikemura, 1985). If we imagine a hypothetical case where a gene is introduced into a new environment (a real-life example could be transfection, horizontal gene transfer, or a viral infection), there is a chance that the CUPrefs are not matching those of the available tRNA pool. Therefore we will observe cis-effects :- i.e. the effect of CUPrefs of the gene on the same gene - and trans-effects : -e.g. the effects of the CUPrefs of one gene on another via the availability/consumption of shared resources(Frumkin et al., 2018). If this new gene uses rare codons, its expression will be most likely slow and overall its expression levels low, as Ribosomes will take a longer time to attach the corresponding tRNAs and thus potentially also affecting the folding of the protein (cis-effect) (Kim et al., 2015; Liu, 2020; Weinberg et al., 2016). This will cause Ribosomes to be stuck for a prolonged period on the gene. Translation is the most resource demanding step of gene expression, especially because of the metabolic cost of Ribosomes production (Dekel & Alon, 2005). Therefore, to minimize energy expense, cells have a limited number of Ribosomes, which also avoids the formation of ribosome traffic jams (Chu & Von Der Haar, 2012).This leads to a fragile equilibrium, where Ribosome perform in an efficient way, but if some of them are stalling, it may rapidly become a limiting factor for gene expression and may also influence the expression of other genes, as they will have less Ribosomes available (trans-effect). The new protein, because of the slow translation may be misfolded, and result in a not properly functioning protein, that can even lead to diseases in some case (Allan Drummond & Wilke, 2009). In this case the cell will either degrade the misfolded protein, or invest in the reparations of the molecule, by chaperon proteins(Walsh, Bowman, Soto Santarriaga, Rodriguez, & Clark, 2020).

A beautiful demonstration of localized CUPref variations is the “ramp” described by Tuller and coworkers at the beginning of mRNAs with undermatching CUPrefs that may serve to slow down the Ribosomes in order to avoid a ribosome traffic jam and therefore, to limit the cost of expression (Tuller et al., 2010). Meanwhile at the end of Eukaryotic mRNAs Tuller and coworkers found an accumulation of matching codons, that would accelerate translation, and hence the detachment of Ribosomes. Mind, that the exact roles of these local CUPrefs variations in a gene, are still under debate.

The two main forces that may explain codon usage pattern are mutation and selection(Duret, 2002; Plotkin & Kudla, 2011). Selection, or more precisely translational selection, posits that synonymous mutations changing CUPrefs have an effect on cellular fitness, and therefore can be selected for or against. This can be linked to altered protein synthesis levels, disregulating the protein cellular

composition, to increased metabolic cost of over-expressing a superfluous gene, or to increased energetic cost of degradation of improperly synthesized/folded proteins (Mordstein et al., 2021). If these scenarios modifying cellular fitness can modify the organism fitness, natural selection could act selecting for or against certain synonymous mutations. As demonstrated among others by Lebeuf-Taylor and coworkers, Distribution of Fitness effect (DFE) of synonymous mutations can be highly variable, ranging from deleterious to beneficial mutations, and thus translational selection does take place (Agashe et al., 2016; Bailey, Hinz, & Kassen, 2014; Lebeuf-Taylor, McCloskey, Bailey, Hinz, & Kassen, 2019).

On the other hand the mutational explanation implies that CUPrefs are a result of other fundamental phenomenons that take place in a cell, for example, biased DNA repair(Kaufmann & Paules, 1996; Lujan et al., 2012), replication(Wolfe, Sharp, & Li, 1989), GC biased gene conversion in vertebrates (Pouyet, Mouchiroud, Duret, & Sémon, 2017)or recombination(Eyre-Walker, 1993). In each of these mechanisms some nucleotides are preferred over others eventually changing (or maintaining) CUPrefs over time.

Ever since the study of CUPrefs started, different ways to quantify it have been developed. The most commonly used are the Codon Adaptation Index (CAI) and the Effective Number Codons (ENC). The CAI uses a reference set of highly expressed genes from a species to assess the relative merits of each codon, and a score for a gene is calculated from the frequency of use of all codons in that gene (Sharp & Li, 1987). Meanwhile the ENC is not dependent on a reference, but quantifies instead how far the codon usage of a gene departs from equal usage of synonymous codons and of only using one codon per amino acid in the sequence (Wright, 1990).

Here in this work we use the Codon Usage Similarity Index (COUSIN) both as method and tool (Bourret, Alizon, & Bravo, 2019). COUSIN was recently developed by members of our team, and has proven to be a reliable and easy to use measure in past and currently running projects. In fact, COUSIN compares the codon usage of the query sequence both to a reference (the host of a pathogen for example) and to a null hypothesis (equal usage of synonymous codons). A score between zero and one, can be interpreted as “matching the reference”, while a score higher than one or lower than zero is “overmatching” or “undermatching” respectively.

Viruses are an especially interesting model to study if we are interested in CUPrefs. As they use the translation machinery of the host cell, they are entirely dependent on their host’s tRNA supply and Ribosomes in most cases. Therefore a matching CUPrefs between the viral genes and the cellular machinery could result in a fast production of great quantity of viral proteins, e.g. of virions. The downside of this, is that the immune system of the vertebrate host, may quickly detect the infection and react to it. This is why, in theory a virus with undermatching CUPrefs would, of course produce

less viral proteins in a more slower pace, but it could avoid rapid elimination by the immune system. Indeed, viral CUPrefs seem to be shaped by many different factors such as : host type, capsid shape, RNA or DNA virus, immune evasion, and the place of transcription and translation (Mordstein et al., 2021).

It is unavoidable to investigate the underlying mechanisms that shape the molecular evolution of pathogens and viruses, in order to control the diseases that impact our world. This could not be more important than right now, in 2021 when the world had to face a virus induced pandemic. Although this work is not exclusively targeting viruses, it was done with the thought of helping understand the way they function and evolve with an emphasis on CUPrefs in eukaryotes. For this, in the next chapter, I will present a study of recombinant papillomaviruses infecting cetaceans with an eye on their regulatory motifs and of course, their codon usage. In chapter 3, I will address the evolution CUPrefs of paralogous genes in vertebrates. And finally in Chapter 4, I will present the results of a long term experimental evolution study where we analyze the effects of CUPrefs at each step of protein synthesis.

Chapter 2

Papillomaviruses infecting cetaceans exhibit signs of genome adaptation following a recombination event

The article that follows, appeared in the sixth volume of the journal *Virus Evolution*, in 2020. It was started as an internship work under the supervision of co-authors Anouk Willemsen and Ignacio G. Bravo, and finished as part of the PhD course.

In this study, I present how a recombination event was followed by evolution in the genome of Papillomaviruses (PV) infecting Cetartiodactyles, with an emphasis on the study of the conserved motifs that regulate gene expression, and on Codon Usage Preferences (CUPrefs) of the viral genes.

As Viruses use the host's cell machinery and resources to express their genes, proteins needed in high quantities often have CUPrefs matching the host's tRNA pool. Human PVs however show CUPrefs that go against that of host cells, but can be linked to their clinical manifestations (Félez-Sánchez et al., 2015).

Papillomaviruses are small non-enveloped dsDNA viruses. They contain three main genomic regions: the Upstream regulatory region (URR), the Early region (E) and the Late region (L). Although recombination has been documented on several occasions in PVs infecting humans (Angulo & Carvajal-Rodríguez, 2007; Bravo & Alonso, 2004; Narechania, Chen, DeSalle, & Burk, 2005), in most cases it happens between closely related strains. Meanwhile a series of PVs infecting exclusively Cetaceans has been found recombinant between two distant crowngroups of PVs(Gottschling et al., 2011; Rector et al., 2008; Robles-Sikisaka et al., 2012).

Int his study we aimed at assessing the impact of recombination on the viral genome, by comparing recombinant PVs to other PVs infecting the same host order : *Cetartiodactyla*, including even-toed ungulates and cetaceans. We looked at their distribution of regulatory motifs, CUPrefs and phylogeny and the relation of clinical traits to these aspects.

We collected all PV genomes infecting Cetartiodactyles at the time of analysis (2018) from GenBank, with their corresponding metadata about host species, anatomical location and clinical presentation. Our dataset regroups 58 PVs from three different Crowngroups, that infect 20 different host species. Eleven of the studied PVs are recombinant, all of them sampled from Cetacean hosts. The particularity of these recombinant PVs, is that their Early region belongs together with the Alpha-Omikron crown group while the late gene region belongs together with that of the Beta-Xi crown group.

After analysis of regulatory motifs, CUPrefs and phylogenetic reconstruction, we propose that a single recombination event lies at the origin of this recombinant group, which allowed them to evolve a set of unique features subsequently. The results suggest that the regulatory motifs identified are not a combination of those in the parental lineages but specific to the recombinant group. The most basal recombinant PV doesn't display these conserved motifs, which further suggest that they appeared as an adaptation to new conditions. Furthermore, the distribution of regulatory motifs correlates well with not just the viral taxonomy but also with that of the host species. This could be interpreted as a regulatory adaptation to the host.

When analyzing the CUPrefs of *Cetartiodactyla* PVs, we observed no significant differences between recombinant and non-recombinant PVs, rather we found that CUPrefs are dependent of gene type and therefore the moment of expression during the natural history of the infection. Indeed, Early and Late genes have easily distinguishable CUPrefs, which is might be to match the different life stages of a Keratinocyte where they are expressed. We also compared the CUPrefs of PVs to that of their hosts by the COUSIN tool. we observed, that, as is the case in other PVs, the studied viruses do not match the CUPrefs of their respective hosts, and displayed instead opposite CUPrefs compatible with an immune system evasion strategy. This hypothesis is consistent with observations in other viruses, where CUPrefs of their genes seem to be an adaptation to the host environment by avoiding immune response (Bahir, Fromer, Prat, & Linial, 2009; Lin et al., 2018).

Virus Evolution, 2020, 6(1): veaa038

doi: 10.1093/ve/veaa038
Research Article

Papillomaviruses infecting cetaceans exhibit signs of genome adaptation following a recombination event

Fanni Borvető,[‡] Ignacio G. Bravo[§], and Anouk Willemsen[*,†,**]

Centre National de la Recherche Scientifique (CNRS), Laboratory MIVEGEC (CNRS IRD Univ, Montpellier), 911 Avenue Agropolis, BP 64501, 34394 Montpellier, France

Corresponding author: E-mail: anouk.willemsen@univie.ac.at

[†]Present address: University of Vienna, Centre for Microbiology and Environmental Systems Science, Division of Microbial Ecology, Vienna, Austria.

[‡]https://orcid.org/0000-0002-2532-7160

[**]https://orcid.org/0000-0002-8511-3244

[§]https://orcid.org/0000-0003-3389-3389

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

Abstract

Papillomaviruses (PVs) have evolved through a complex evolutionary scenario where virus–host co-evolution alone is not enough to explain the phenotypic and genotypic PV diversity observed today. Other evolutionary processes, such as host switch and recombination, also appear to play an important role in PV evolution. In this study, we have examined the genomic impact of a recombination event between distantly related PVs infecting Cetartiodactyla (even-toed ungulates and cetaceans). Our phylogenetic analyses suggest that one single recombination was responsible for the generation of extant 'chimeric' PV genomes infecting cetaceans. By correlating the phylogenetic relationships to the genomic content, we observed important differences between the recombinant and non-recombinant cetartiodactyle PV genomes. Notably, recombinant PVs contain a unique set of conserved motifs in the upstream regulatory region (URR). We interpret these regulatory changes as an adaptive response to drastic changes in the PV genome. In terms of codon usage preferences (CUPrefs), we did not detect any particular differences between orthologous open reading frames in recombinant and non-recombinant PVs. Instead, our results are in line with previous observations suggesting that CUPrefs in PVs are rather linked to gene expression patterns as well as to gene function. We show that the non-coding URR of PVs infecting cetaceans, the central regulatory element in these viruses, exhibits signs of adaptation following a recombination event. Our results suggest that also in PVs, the evolution of gene regulation can play an important role in speciation and adaptation to novel environments.

Key words: virus evolution; recombination; gene regulation; papillomavirus.

1. Introduction

Papillomaviruses (PVs) are small, non-enveloped viruses with double-stranded DNA genomes varying between 5.7 and 8.6 kb in size. The minimal PV genome consists of an upstream regulatory region (URR), an early gene region encoding the *E1* and *E2* genes, and a late gene region encoding the *L2* and *L1* genes. Other genes that are not strictly conserved in all PV genomes

© The Author(s) 2020. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

are *E4* (nested within *E2*), *E5*, *E6*, *E7*, and *E10*. As the names suggest, the early genes are expressed during the early stages of PV infection, while the capsid proteins L2 and L1 are expressed during later stages.

According to the International Committee on Taxonomy of Viruses (https://talk.ictvonline.org/taxonomy/), the *Papillomaviridae* family currently consists of >50 genera and >130 species (Van Doorslaer et al. 2018). Based on the phylogenetic relationships of the concatenated early and late core genes (*E1-E2-L2-L1*) PVs have been classified into a limited number of crown groups: Alpha-Omikron, Beta-Xi, Lambda-Mu, and Delta-Zeta (Gottschling et al. 2011b; Bravo and Felez-Sanchez 2015). PVs have a wide host range, infecting bony fishes, birds, reptiles, and virtually all mammals (Antonsson and Hansson 2002; Rector and Van Ranst 2013; López-Bueno et al. 2016). However, the best-known members of the *Papillomaviridae* are PVs infecting humans, because of the clinical importance of some of these infections.

Although PVs have evolved in close relationship with their hosts, virus–host co-evolution alone is not enough to explain the phenotypic and genotypic viral diversity observed today (Bravo and Félez-Sánchez 2015; Gottschling et al. 2011b). Other processes such as host switch and recombination also play an important role in PV evolution (Rector et al. 2008; Gottschling et al. 2011b). Recombination remains a rare event for PVs, because even if individual mammals are very often infected by several different PVs at any given time, recombination requires the simultaneous presence of two different PV genomes within the same infected cell. Nevertheless, the result of a recombination event is most often conspicuous, rendering a chimeric daughter genome easily identifiable because of their differential similarities with the parental ones along the sequence. Evidence of recombination has been described within the group of PVs infecting Primates that includes the most oncogenic PVs to humans (Bravo and Alonso 2004; Narechania et al. 2005; Angulo and Carvajal-Rodríguez 2007). Another compelling example of recombination between distant viral sequences are two viruses isolated from bandicoots, where the early gene region resembles those of Polyomaviruses and the late gene region resembles those of PVs (Woolford et al. 2007; Bennett et al. 2008). However, the most noticeable lineage of recombinant PVs is a group of viral genomes isolated from different cetacean species (whales, dolphins, and porpoises), with the early gene region resembling that of PVs in the Alpha-Omikron crown group and the late gene region resembling that of PVs in the Beta-Xi crown group (Rector et al. 2008; Gottschling et al. 2011a; Robles-Sikisaka et al. 2012).

Recombination events between distantly related viruses can lead to drastic genomic changes. For example, a recombination event may change the repertoire of genes present in the genome, or modify the match between the codon usage preferences (CUPrefs) of virus and host. As a consequence, upon recombination adaptive changes may occur in both coding and non-coding regions of the viral genome. For the non-coding regions, sequence changes may occur in regulatory sites. For PVs, regulatory elements are mainly found in the URR, which contains transcription-factor binding sites (TFBSs) and other regulatory motifs that are necessary to regulate replication and transcription of the virus, with viral E1 and E2 as the central interaction partners (Bernard 2013). As an ATP-dependent DNA helicase, the PV E1 protein is essential for replication and amplification of the viral episome. Viral DNA replication is initiated by E1 binding to specific sequence motifs, such as the palindromic AT-rich E1-binding site (E1BS) and other versions of E1BSs, located within the URR (Bergvall, Melendy, and Archambault 2013). These E1BSs are often regarded as the 'origin of DNA replication'. The E2 viral protein is an essential transcription regulator that binds specifically to 12 bp motifs—E2-binding sites (E2BSs)—located mostly within the URR (McBride 2013). In addition, E2 modulates the shift from early to late transcript production, acting independently on E2BS outside the URR (Johansson et al. 2012).

In the coding regions, the CUPrefs of a virus may shift after drastic genomic changes. Since PVs depend on the host translation machinery and on the available host tRNAs, one can expect that viruses would evolve to match their CUPrefs to those of the host. Therefore, proteins required in large amounts are usually encoded by genes optimized to the host's CUPrefs while a poor match of CUPrefs generally results in lower protein production (Bahir et al. 2009). Despite this observation, it has been shown that CUPrefs of human PVs do not match those of their host, but can instead be associated to different clinical presentations of the infections; viruses causing productive lesions display CUPrefs closer to those of the host than viruses causing more oncogenic lesions (Félez-Sánchez et al. 2015). In addition, the timing of expression—early gene expression in basal epithelium versus late gene expression in differentiating epithelium—largely determines the differential CUPrefs (Zhou et al. 1999; Félez-Sánchez et al. 2015).

In this study, we have examined the recombinant cetacean PVs as well as closely related PVs infecting Cetartiodactyla (even-toed ungulates and cetaceans). To better understand what drives the evolution of these viruses, we have correlated their genomic content to their phylogenetic relationships. In particular, we have investigated whether recombinant PVs contain unique regulatory motifs and whether the recombinant and non-recombinant PVs are different in their CUPrefs. In addition, we have analysed whether viral CUPrefs are similar to those of the hosts they infect and whether macroscopic traits of the corresponding infection (e.g. clinical presentation or anatomical site of the infection) correlate with CUPrefs, motif distribution, and phylogenetic clustering. These tests allowed us to investigate the impact of recombination on the genomes of PVs infecting Cetartiodactyla.

2. Materials and methods

2.1 PV genome sequences and their characteristics

We collected the complete genomes of PVs infecting Cetartiodactyla from the Papillomavirus Episteme database (PaVE: https://pave.niaid.nih.gov/) and GenBank (https://www.ncbi.nlm.nih.gov/genbank/) between March and May 2018. The ORFs (*E10*, *E6*, *E7*, *E1*, *E2*, *E4*, *E5*, *L2*, and *L1*) and the URR of 58 reference PV genomes were extracted for subsequent analyses. For each PV genome, we collected information on the corresponding host species, the clinical presentation of the infection, the anatomical location, and viral taxonomy, as reported by the authors in the corresponding PaVE and GenBank entries or publications (Supplementary Table S1). The PVs in this study were sampled from twenty different host species that belong to seven distinct host families (Bovidae, $n = 30$; Camelidae, $n = 2$; Cervidae, $n = 11$; Delphinidae, $n = 8$; Giraffidae, $n = 1$; Phocoenidae, $n = 4$; Suidae, $n = 2$). They represent three viral crown groups: Alpha-Omikron ($n = 13$), Beta-Xi ($n = 18$), and Delta-Zeta ($n = 20$), along with several unclassified viral genomes ($n = 7$). Most of the viral genomes have been retrieved from benign epithelial lesions ($n = 47$), albeit a number of

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

samples correspond to malignant lesions ($n = 3$), asymptomatic infections ($n = 7$), and one eye fluid sample. The data set contains eleven recombinant PV genotypes infecting members of the Delphinidae and Phocoenidae host families. These genomes have already been identified as being recombinant by previous studies (Rector et al. 2008; Gottschling et al. 2011a; Robles-Sikisaka et al. 2012).

2.2 Phylogenetic inference

For the construction of phylogenetic trees, we first used the concatenated *E1-E2* genes and the concatenated *L2-L1* genes. Two different data sets were used, one including all PVs collected for this study, and a second one removing the recombinant PVs. The individual gene sequences were aligned at the amino acid level using MUSCLE in Geneious v8.0.5 (https://www.geneious.com/), and subsequently back-translated to nucleotides. The nucleotide alignments were filtered with Gblocks v.0.91b (Castresana 2000) to exclude the non-informative positions. The Gblocks parameters used were as follows: type of sequence: codons; minimum number of sequences for a conserved position: thirty; minimum number of sequences for a flank position: thirty; maximum number of contiguous non-conserved positions: twelve; minimum length of a block: six; allowed gap positions: all; use similarity matrices: yes. The phylogenies of the concatenated *E1-E2* and *L2-L1* alignments were used to construct maximum likelihood (ML) trees. ML phylogenetic inference was done at the nucleotide and amino acid level with RAxML 8.2.9 (Stamatakis 2014), under the GTR + Γ4 model, using 5,000 bootstrap cycles and three partitions (one for each codon position). Additional ML trees were constructed (GTR + Γ4 model, 10,000 bootstrap cycles, one partition per codon position) for each of the individual *E1*, *E2*, *L1*, and *L2* genes that were used for comparing the phylogenetic signal with the CUPrefs.

2.3 Comparison of early gene and late gene phylogenetic trees

To measure topological distances between the early (*E1-E2*) and late (*L2-L1*) gene trees, we compared pairwise distances, the Robinson–Foulds (RF) (Robinson and Foulds, 1981) distance, and the K-tree score, using Ktreedist v.1.0 (Soria-Carrasco et al. 2007). The calculated pairwise distances in the two corresponding trees were compared by a Mantel test, to evaluate whether correlation between the two matrices was higher than expected by chance. The RF distance evaluates the differences between two trees by counting the number of partitions that are not present in both trees. The maximum RF distance is thus the total number of nodes in both trees and would correspond to two trees that do not share any partition. The K-tree score is the minimum branch length distance one can get from one tree to another after scaling one of them. The higher the RF distance and K-tree score, the bigger the topological dissimilarity between the two trees. The tree distance measures were calculated between nucleotide-based trees, amino acid-based trees, and between trees with and without recombinant taxa.

2.4 Distribution of conserved motifs in the upstream regulatory region

We used the MEME Suite v.4.11.0 (Bailey et al., 2009) to identify conserved motifs in the URR. Some of the PV genomes studied here contain a very short URR that is followed by the *E10* ORF. For these PVs, we concatenated the URR and *E10* for the analysis, as we suspect that *E10* may be functionally linked to the short URRs. We scanned for motifs on both strands of the URR, with a length between six and fifty nucleotides, and with a minimum of four occurrences in total per motif. To determine the *E*-value cut-off ($E = 3.63 \times 10^6$) for the discovered motifs, we shuffled each of the sequences from the same data set (conserving the sequence length and nucleotide composition), and repeated the analyses. We constructed a matrix containing the absolute counts of the detected motifs and analysed this matrix by a centred principal component analysis (PCA), and a correspondence analysis (COA). The detected motifs were also compared with the known regulatory motifs in PVs (Bergvall, Melendy, and Archambault 2013; Bernard 2013; McBride 2013), as well as with those in the online databases TOMTOM (Gupta et al. 2007) and TRANSFAC (Wingender 2000). For certain PV genomes for which important motifs were not detected in the URR, we used FIMO implemented in the MEME Suite to scan for the presence of these motifs elsewhere in the genome.

2.5 Codon usage preferences

We calculated the CUPrefs for all ORFs of the fifty-eight PV genomes included in this study. The relative frequencies for each of the eighteen families of synonymous codons were calculated using COUSIN v.1.0 (Bourret et al. 2019). We only considered the frequencies of the fifty-nine codons with redundancy (i.e. excluding Met, Trp, and stop codons). A matrix was created in which the rows correspond to the ORFs and the columns to the fifty-nine relative frequency values, such that each row contains the codon usage information for a specific ORF. We performed a PCA to display the variance distribution and dispersion of CUPrefs for orthologous ORFs as well as for all ORFs present within the same genome.

In addition, we used COUSIN to compare the viral CUPrefs to those of the corresponding host species. The algorithm in this program allows us to compare the CUPrefs of a query (ORFs of PV genomes) to those of a reference data set (ORFs of host genomes) and outputs a normalized value. The COUSIN score can be interpreted as follows: COUSIN = 1, the CUPrefs of the PV ORFs are similar to those of the corresponding host; COUSIN = 0, the CUPrefs of the PV ORFs are similar to a random usage of synonymous codons; COUSIN < 0, the CUPrefs of the PV ORFs are opposite to those of the corresponding host (i.e. the less used codons in the host reference are used more often in the query than in the null hypothesis of equal frequency), and COUSIN > 1, the CUPrefs of the PV ORFs are superior to those in the reference (i.e. the more frequent codons in the host reference are even more frequently used in the query) (Bourret et al. 2019). To calculate the CUPrefs of the hosts, a representative genome for each host family was chosen and the respective codon usage tables were calculated. The representatives used are: Bovidae—*Bos Taurus* (accession: AC_000158), Camelidae—*Camelus dromedarius* (accession: NW_011590949), Cervidae—*Odocoileus virginianus* (accession: NW_018326927), and Suidae—*Sus scrofa* (accession: NC_010443). For PVs infecting Delphinidae and Phocoenidae, we chose a common representative, *Tursiops truncatus* (accession: NW_017842062), as both host families are closely related. We did not calculate the CUPrefs for the Giraffidae family as the available giraffe genomes (GenBank accessions: LVKQ00000000.1 and SJXV00000000.1) are not annotated, hence GcPV1 was removed from the COUSIN analysis.

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

2.6 Statistical analyses

Statistical analyses and graphics were done using R v.3.4.3 (R Core Team 2018), with the aid of the packages 'ape' and 'vegan'. To compare the phylogenetic trees, we calculated pairwise distances between the concatenated *E1-E2* and *L2-L1* trees and between all single gene trees (*E1*, *E2*, *L2*, and *L1*). Jaccard distances were calculated for the distribution of motifs, and Euclidian distances for the CUPrefs of the different genes. Correlation between distance matrices were then evaluated with a Mantel test. To investigate whether the viral taxonomy, host taxonomy, sampling location, and clinical presentation correlate with the CUPrefs of all PV ORFs, the phylogenetic signal of the early gene and late gene trees, and the distribution of motifs, we used a permutational multivariate analysis of variance (PERMANOVA).

3. Results

3.1 Phylogenetic reconstruction of PVs infecting Cetartiodactyla

We collected 58 PVs infecting Cetartiodactyla from the PaVE and GenBank databases (Supplementary Table S1). ML phylogenetic trees of the concatenated early genes (*E1-E2*) and the concatenated late genes (*L2-L1*) were constructed at the nucleotide (Fig. 1) and amino acid (Supplementary Fig. S1) levels. The constructed trees are well supported with high bootstrap values, although few inner branches have low (>30 and <50) bootstrap values.

In both *E1-E2* and *L2-L1* phylogenetic trees at nucleotide and amino acid levels, the Delta-Zeta crown group (blue in Fig. 1 and Supplementary Fig. S1) forms a monophyletic clade. The other crown groups, Alpha-Omikron (coloured orange/red) and Beta-Xi (coloured green), and unclassified PVs (coloured purple), form monophyletic clades in the early gene trees. However, these do not appear to be monophyletic in the late gene trees (Fig. 1 and Supplementary Fig. S1). This incongruence is due to the 'chimeric' genomic composition of the recombinant cetacean PVs, and thereby, a position in the phylogenetic trees that varies depending on the genome region considered. In the early gene tree, these recombinant PVs (in red) cluster with non-recombinant Alpha-OmikronPVs (PphPV4, SsPV1, in orange), while in the late gene tree, the recombinant PVs cluster with Beta-XiPVs (in green) infecting Bovidae and Cervidae. Despite this displacement and several internal changes (as shown with the tanglegram in Fig. 1 and Supplementary Fig. S1), recombinant PVs remain monophyletic in both trees, suggesting that only one main recombination event occurred in the ancestral genome of these PVs.

To measure topological distances between the constructed phylogenetic trees, we calculated the pairwise distances, the K-tree scores, and the RF distances (Table 1 and Supplementary Table S2). The pairwise distances were compared with a Mantel test, a statistical test indicating correlation between the two matrices. We first compared the distances between all amino acid and nucleotide-based *E1-E2* trees and did the same for the *L2-L1* trees. None of the three distance measures indicate a significant difference between the amino acid and their corresponding nucleotide-based phylogenetic trees (early vs. early and late vs. late in Supplementary Table S2). Upon comparing the early and late gene trees without the recombinant strains, we also observe a high correlation (>0.95) between trees (Table 1 and Supplementary Table S2). However, upon

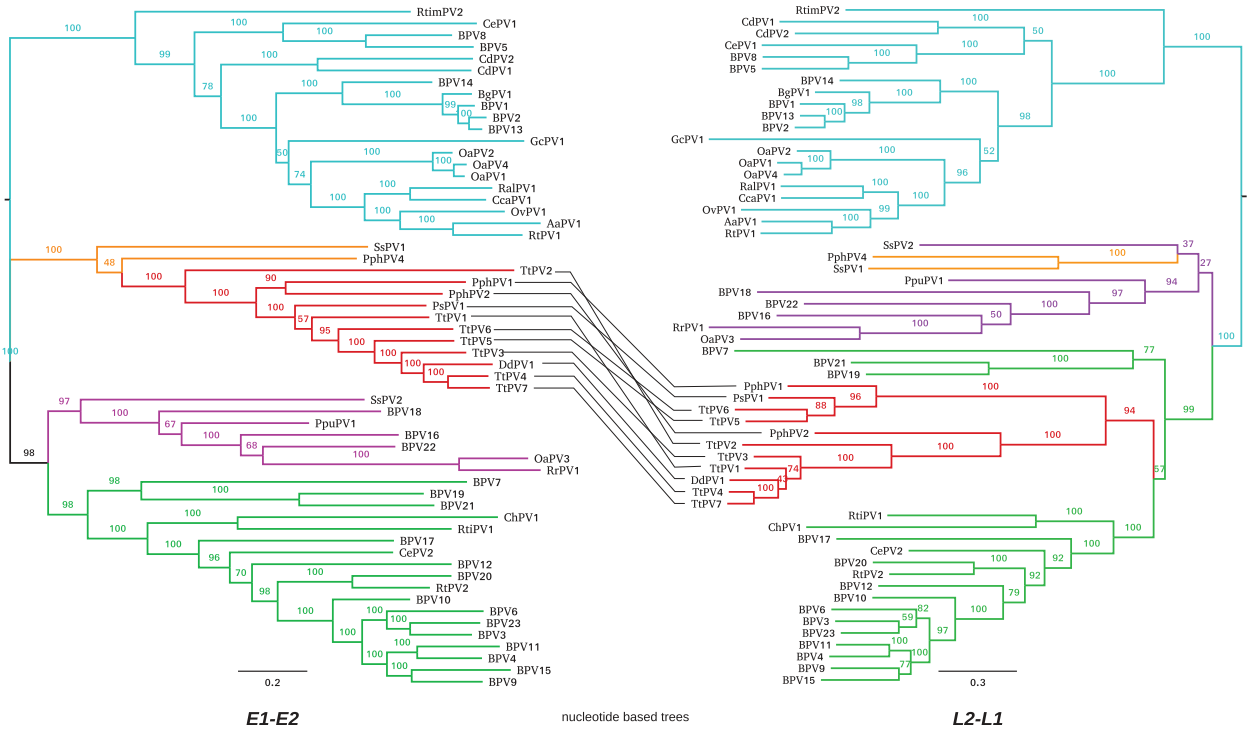


Figure 1. ML nucleotide-based phylogenetic trees of the concatenated *E1-E2* (early) and *L2-L1* (late) gene alignments. Both trees comprise fifty-eight PVs infecting Cetartiodactyla. The colour code highlights the different PV clades based on the PV crown groups: orange, Alpha-OmikronPVs; red, recombinant PVs clustering with the Alpha-OmikronPVs in the *E1-E2* tree; green, Beta-XiPVs; blue, Delta-ZetaPVs; and purple, yet unclassified PVs. Values at the branches correspond to bootstrap support values. A tanglegram connects the recombinant cetacean PVs between the early and late gene trees, emphasizing the differences in positioning of these PVs.

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

Table 1. Distances between phylogenetic trees based on the early and late gene regions.

Trees compared	Mantel test correlation	*P*-value	K-tree score	RF distance
E1-E2 nt–*L2-L1* nt	0.9577	<0.001	0.9330	12
E1-E2[a] nt–*L2-L1*[a] nt	0.8824	<0.001	1.4766	32
E1-E2 aa–L2-L1 aa	0.9506	<0.001	1.1680	10
E1-E2[a] aa–L2-L1[a] aa	0.8745	<0.001	1.9529	28

The nucleotide and amino acid-based E1-E2 and L2-L1 trees are compared by using pairwise distances and a subsequent Mantel test with the corresponding *P* values, by the K-tree score, and by the RF distances. The introduction of the recombinant taxa in the phylogenetic inference is accompanied by a loss in concordance between the phylogenetic reconstructions for early and late genes.
nt, nucleotide-based tree; aa, amino acid-based tree.
aTree includes recombinant taxa.

introducing the recombinant taxa, this correlation is lower (~0.88). In concordance with the Mantel test, the K-tree scores and RF distances are higher for comparisons of trees that include the recombinant PVs, indicating that the number of topological incongruences is higher.

3.2 The distribution of conserved motifs in the URR reflects the phylogenetic relationships

The URR in PV genomes harbours TFBSs and other conserved motifs that regulate viral replication and transcription. The number and occurrence of these conserved motifs are more important than their order of appearance. To investigate whether the recombination event led to changes in the presence/absence of regulatory motifs and therewith possible changes in PV replication, we scanned for conserved motifs in the URR of the PV genomes. The MEME algorithm detected twenty-two conserved motifs throughout the URR of the fifty-eight query sequences (Fig. 2 and Supplementary Fig. S2). The most recurrent ones were identified as E2BSs (M1 in Fig. 2 and Supplementary Fig. S2) and the preferred E1BS (M2 in Fig. 2 and Supplementary Fig. S2). The E2BS was detected 298 times in 56 out of 58 sequences, as we could not detect this motif in the URR of BPV19 and BPV21. The E1BS was detected sixty times in all fifty-eight sequences. Since the E1- and E2BSs are pivotal for the PV life cycle, we suspect that the E2BS is located elsewhere in the genome of PVs lacking these in the URR. Indeed, for both BPV19 and BPV21, we detected an E2BS within the *L2* gene. Moreover, it is likely that additional E1- and/or E2BSs were not detected due to sequence divergence from the consensus motif sequence of PVs included in this study. Apart from the E2BS and E1BS, we detected twenty other motifs. However, we were not able to match these motifs with other known PV regulatory motifs or with those known in the online databases (TOMTOM and TRANSFAC). Certain of the URR motifs seemed to be exclusive to specific PVs; motifs M8, M9, M10, and M15 are present only in recombinant PVs, while motif M6 is solely found in a smaller group of Beta-XiPVs (Fig. 2 and Supplementary Fig. S2). The conservation of these motifs in these phylogenetic clades indicates a genuine role for the life cycle of these PVs.

To evaluate the match between the phylogenetic signal and the distribution of detected motifs, we calculated Jaccard distances on the presence/absence matrix of motifs, and compared these to the pairwise distances calculated on the early and the late gene trees. The results show that there is a correlation of 47.8 per cent ($P < 0.001$) between the distribution of motifs and the early gene phylogeny, and a 35.6 per cent ($P < 0.001$) correlation between motifs and the late gene phylogeny.

To analyse the motif distribution in the URR of the different PV genomes, we performed a centred PCA (Fig. 3a). The first axis explains 26 per cent of the observed variance and clearly separates the recombinant, Alpha-Omikron, Beta-Xi, and unclassified PVs from most PVs in the Delta-Zeta crown group. The second axis, explaining 17 per cent of the variance, separates the recombinant PVs (except one) and certain Delta-ZetaPVs from the Beta-Xi and unclassified PVs. More importantly, ten out of the eleven recombinant PVs (in red) are clearly separated from the non-recombinant Alpha-OmikronPVs (in orange). The one exception is a recombinant PV isolated from a bottlenose dolphin (TtPV2), that surprisingly does not cluster with the other recombinant PVs, including six other TtPVs. We relate this observation to the lack of sequence motifs M8, M9, M10, and M15 in the URR of TtPV2 (Fig. 2), which are conserved in and exclusive to all other recombinant PV genomes. In addition to a centred PCA, we also performed a COA to analyse the proportions between the motifs detected (Fig. 3b). The results are highly similar as those obtained for the PCA, where the recombinant PVs are separated from the non-recombinant PVs. The non-recombinant Alpha-OmikronPV (PphPV4) that is positioned closest to the recombinant PVs is also the PV with the closest phylogenetic relationship in the early gene tree (Fig. 1). The main difference between the PCA and the COA results is that certain Beta-Xi PVs, that contain motif M6, are separated from all other PVs (including other Beta-Xi PVs), that do not contain motif M6.

3.3 Orthologous Cetartiodactyla PV genes have similar codon usage preferences

To test whether the CUPrefs of the genes in the recombinant PV genomes are similar to those in the other Cetartiodactyla PV genomes, we calculated the relative frequencies of the fifty-nine codons in synonymous families and displayed this multidimensional information using a PCA. When including all ORFs in the analysis (*E1*, *E2*, *E4*, *E5*, *E6*, *E7*, *E10*, *L2*, and *L1*), we observe that the first axis (explaining 14% of the variance) separates the *E4* ORFs from the rest (Supplementary Fig. S3). This PCA also separates *E10* of BPV4, BPV9, BPV12, BPV15, and BPV23 from the rest. The centre of the PCA contains the *E1*, *E2*, *L2* and *L1* 'core' genes, indicating that these display similar CUPrefs. Although the CUPrefs of *E6* and *E7* do not display a clear pattern, these ORFs cluster closer to the core genes than *E4*, *E5*, or *E10* do. Subsequently, we performed a PCA on the CUPrefs of only the core genes (*E1*, *E2*, *L2*, and *L1*; Fig. 4). The first axis captured 16 per cent of the variance and separates the *E1* ORFS from the *E2* ORFs. The second axis contained 8 per cent of the overall variance and roughly separates the early genes (*E1* and *E2*) from the late genes (*L2* and *L1*). Although the CUPrefs of the late genes partially overlap, the recombinant PVs separate clearly from the other PVs and the first axis splits recombinant *L1* from recombinant *L2*. The relatively low median absolute deviation for each of the studied groups indicates that PVs belonging to the same clade tend to have similar CUPrefs. Unexpectedly, we observed that the CUPrefs of SsPV1 (a non-recombinant Alpha-OmikronPV, recovered from pigs) are very different from those of other PVs and SsPV2 (a non-recombinant unclassified PV, also recovered from pigs) (Fig. 4 and Supplementary Fig. S3).

Subsequently, we investigated whether the differential gene CUPrefs are related to phylogenetic clustering and/or to the

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

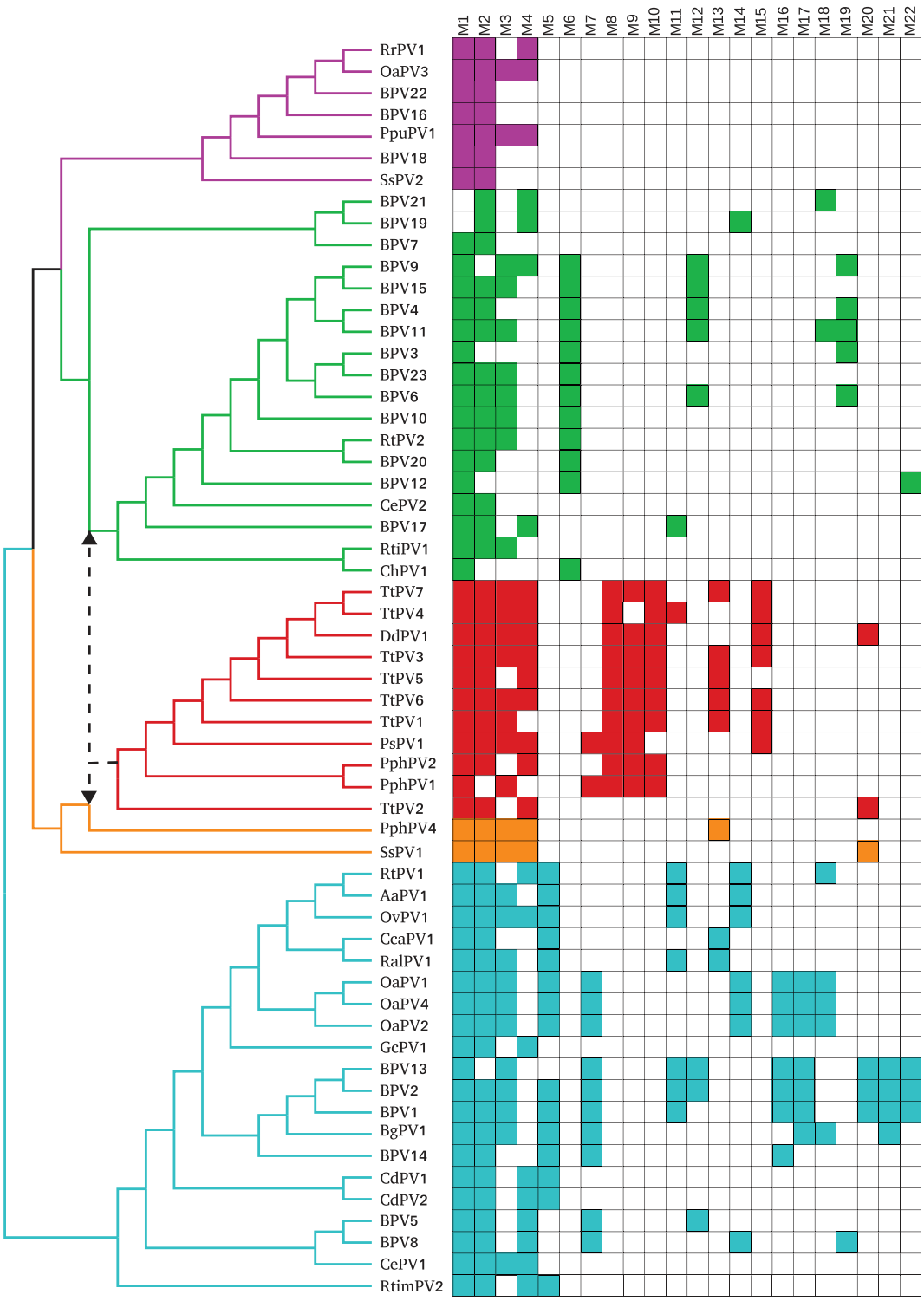


Figure 2. Presence–absence matrix of conserved motifs detected in the URR of fifty-eight PVs infecting Cetartiodactyla. In total, twenty-two motifs were identified by the MEME algorithm, as indicated with M1 to M22 on top of the columns of the matrix. Left to the rows of the matrix, the names of the studied PVs are given and a schematic representation of their phylogenetic relationships is shown. The dashed lines with arrows indicate the different phylogenetic positions of the recombinant PV clade (in red) in the early (placed with PVs in orange) and late gene trees (placed with PVs in green). Colour code corresponds to the different PV clades based on the PV crown groups: orange, Alpha-OmikronPVs; red, recombinant PVs clustering with the Alpha-OmikronPVs in the *E1-E2* tree; green, Beta-XiPVs; blue, Delta-ZetaPVs; and purple, yet unclassified PVs. A filled rectangle means that the given motif was detected in the URR of the given PV. Motifs are numbered and ordered by their abundance. M1 and M2 correspond respectively to the canonical E2BS and E1BS.

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

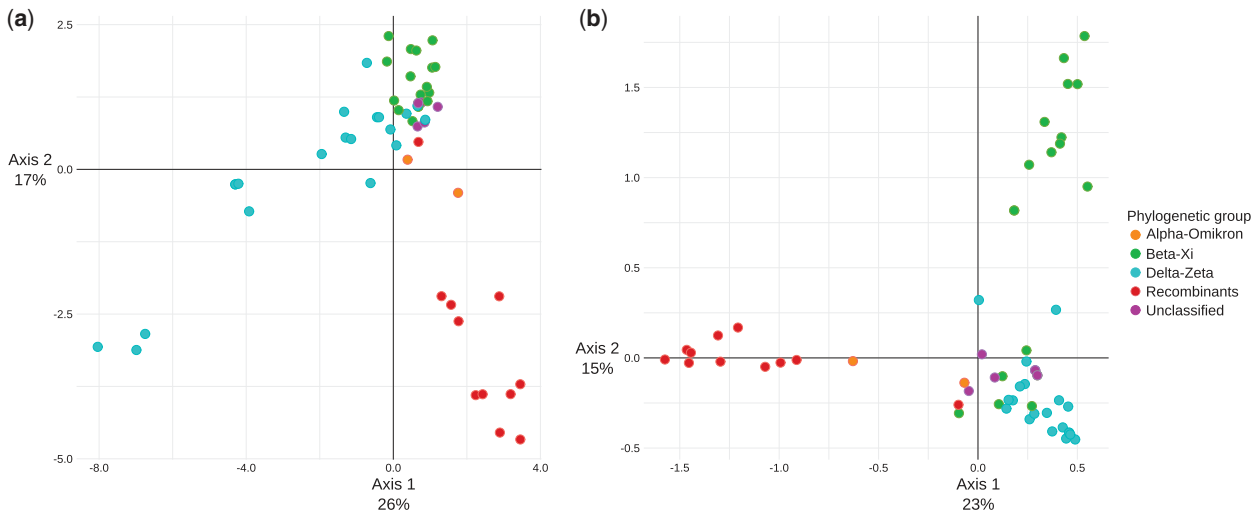


Figure 3. Centred PCA (a) and COA (b) on the distribution of motifs detected in the URR of fifty-eight PVs infecting Cetartiodactyla. As indicated in the legend on the right, colour code corresponds to the different PV clades based on the PV crown groups: orange, Alpha-OmikronPVs; red, recombinant PVs clustering with the Alpha-OmikronPVs in the *E1-E2* tree; green, Beta-XiPVs; blue, Delta-ZetaPVs; and purple, yet unclassified PVs. Values next to the axes represent the percentage of total variance explained by the corresponding axis. For the PCA, the first and second axes represent 43 per cent of the total information. For the COA, the first and second axes represent 38 per cent of the total information.

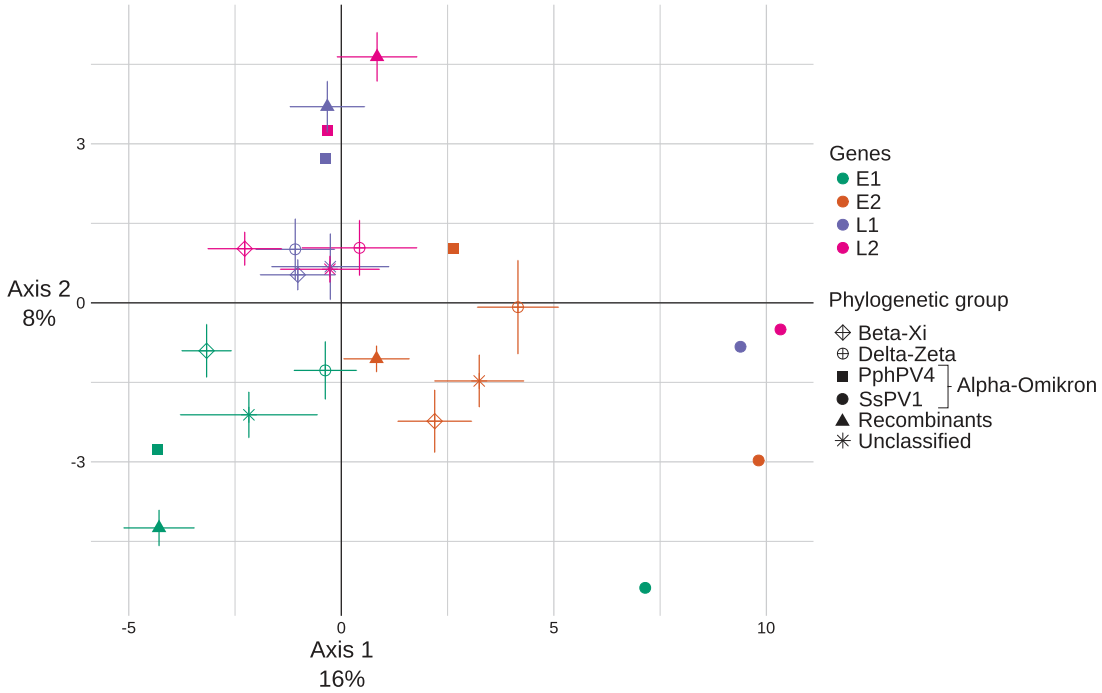


Figure 4. PCA on the CUPrefs of the PV core genes (*E1*, *E2*, *L2* and *L1*) of fifty-eight PVs infecting Cetartiodactyla. The data points are Huber M-estimator values, and the error bars correspond to the median absolute deviation. Colour code corresponds to data stratification by gene. Shapes for data points correspond to data stratification by PV crown group. Values next to the axes represent the percentage of total variance explained by the corresponding axis. Combined, the first and second axes represent 24 per cent of the total information. The main explanatory factor seems to be driven by all genes in SsPV1, infecting pigs. Secondarily, axis 1 splits the early genes *E1* and *E2*, while axis 2 splits the late and the early genes.

presence/absence of motifs in the URR. Therefore, we compared the CUPrefs of the *E1*, *E2*, *L1*, and *L2* genes to the respective gene phylogenetic trees and the URR motif distribution. We observe a higher correlation between CUPrefs and phylogenetic signal than between CUPrefs and motif distribution (Table 2). This is not surprising, as the regulatory motifs analysed in this study are located in a non-coding region, and codon usage is thus not expected to be an important factor in the evolution of this region. Even so, the CUPrefs of the early genes are better correlated to both phylogenetic signal and motif distribution than the CUPrefs of the late genes that show no correlation at all with motif distribution.

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

Table 2. Comparison of the CUPrefs with phylogenetic pairwise distances and URR motif distribution.

	Codon usage versus pairwise phylogenetic distances		Codon usage versus URR motif distribution	
ORF	Mantel test	*P*-value	Mantel test	*P*-value
E1	0.4606	<0.001	0.2246	0.001
E2	0.3424	<0.001	0.2111	0.001
L2	0.1386	<0.001	−0.0443	0.722
L1	0.2555	<0.001	0.0666	0.098

A Mantel test was used to compare the pairwise Euclidian distances of CUPrefs with the corresponding pairwise phylogenetic distances. Similarly, a Mantel test was used to compare the pairwise Euclidian distances of CUPrefs with the corresponding pairwise Jaccard distances of the presence/absence matrix of conserved motifs detected in the URR. This comparison was done for each of the PV core ORFs (*E1*, *E2*, *L2*, and *L1*). Phylogenetic relatedness correlates stronger with CUPrefs for early than for late genes, Similarly, a significant correlation between CUPrefs and the repertoire of motifs in the URR is only observed for the early genes.

3.4 Cetartiodactyla PV codon usage preferences do not follow those of their respective hosts

To investigate whether the Certartiodactyla PVs, and in particular the recombinant cetacean PVs, have similar CUPrefs to the hosts they infect, we calculated the COUSIN score for each of the PV ORFs (see Section 2). As a general observation, for *E6*, *E7*, *E2*, and *E5*, we obtained a COUSIN score close to 0 (Supplementary Fig. S4), indicating that for these ORFs the CUPrefs are not different from a random usage of synonymous codons. The *E4* ORF has a COUSIN score of around 1, significantly higher than all other PV ORFs (Wilcoxon–Mann–Whitney two-sided test: $W = 1{,}7762$, $P < 2.2e\text{-}16$), indicating that the CUPrefs of *E4* are closer to those of the corresponding hosts. The *E10*, *E1*, *L2*, *and L1* genes display COUSIN scores lower than 0 (Wilcoxon–Mann–Whitney one sided test: $V = 1{,}406$, $P < 2.2e\text{-}16$), indicating that the less used codons in the host reference are used more often in the PV ORFs, going towards 'opposite' CUPrefs.

Most of the Certartiodactyla PVs follow the pattern described above, however, after stratifying the COUSIN data per gene and per taxa, we observe individual exceptions (Fig. 5). Contrary to the general observation, the CUPrefs of the *E6*, *L2*, and *L1* ORFs for most recombinant- and closely related non-recombinant PVs in the Alpha-Omikron crown group are closer to those of the hosts as compared to the other PVs. Also, in this phylogenetic group, the CUPrefs of *E4* for three taxa (DdPV1, TtPV4, and TtPV5) display high COUSIN scores (Fig. 5), meaning that the most frequent codons used in the host are even more often used in this ORF. With the PCA in Fig. 4 we already showed that the CUPrefs of SsPV1 are different from those of other PV taxa. With the COUSIN score (Fig. 5), SsPV1 also distinguishes itself from the other Cetartiodactyla PVs. For all ORFs, SsPV1 has CUPrefs close to those of the host (*S.scrofa*).

3.5 Viral taxonomy and host phylogeny explain most of the observed differences in clinical presentation of the infection

PERMANOVA tests were performed to investigate whether qualitative traits—viral taxonomy, host taxonomy, sampling location, and clinical presentation—correlate with CUPrefs, motif distribution, and phylogenetic clustering of PVs. The best correlation was found between the concatenated early (*E1-E2*) and late (*L2-L1*) gene trees and viral taxonomy (60% and 52%, respectively), while for the other traits this correlation was much lower (Table 3). We also observe that the distribution of conserved motifs in the URR correlates best with viral taxonomy (33%), followed by the host taxonomy (26%). For CUPrefs, we observed that all ORFs correlate better with host taxonomy than with viral taxonomy (Table 3). This is an unexpected result as we have shown that the CUPrefs of *E6*, *E7*, *E2*, and *E5* are similar to a random usage of synonymous codons and that the CUPrefs of *E10*, *E1*, *L2*, and *L1* are going towards and 'opposite' direction as compared to the CUPrefs of the hosts they infect (Supplementary Fig. S4). These results suggest that even though the PV CUPrefs do not necessarily match those of the hosts, the viral CUPrefs do seem to be partially modulated by interaction with the different host species. The CUPrefs of *E1* correlate best with host taxonomy (35%), followed by *E5* (29%, Table 3). As *E5* was not included in the CUPrefs analysis in Fig. 4, here we investigated this ORF separately. *E5* is only present in the genomes of PVs belonging to the Delta-ZetaPV crown group, consisting of PVs infecting bovids, cervids, and one giraffid. When performing a PCA on the E5 CUPrefs (Supplementary Fig. S5), the PVs infecting Bovidae are separated by the first axis (explaining 21% of the observed variance) into two clusters, one cluster with PVs infecting members of the Bos genus, and a second cluster with PVs infecting members of the Ovis genus. The giraffid PV (GcPV1) clusters with the Ovis group. Both the first and the second axis (explaining together 36% of the variance), separate PVs infecting Cervidae from the rest. Overall, these results suggest that also for *E5* the CUPrefs are host genus specific.

4. Discussion

Here we analysed PVs infecting Certartiodactyla with the main aim to better understand the evolution of recombinant PVs infecting cetaceans. Discrepancies between the early and late gene trees are compatible with a recombination event between ancestral PVs belonging to two distant viral clades, with extant descents classified today into two different crown groups (Alpha-Omikron and Beta-Xi PVs) (Gottschling et al. 2011a; Robles-Sikisaka et al. 2012). Our phylogenetic analyses suggest that one single recombination event occurred between the genomes of these distantly related PVs. Our results for the phylogenetic inference are consistent with those communicated for the complete viral family, with recombinant cetacean PVs clustering with non-recombinant cetacean PVs in the Alpha-Omikron crown group in the early gene tree and as a sister clade to the XiPVs in the Beta-Xi crown group in the late gene tree (Supplementary material in Willemsen and Bravo 2020). The ancestral genomes of these two clades were dated back to around 60 and 70 million years ago (Ma), respectively (Willemsen and Bravo 2020), suggesting that the recombination event occurred between 60 Ma and the present.

Our analyses here presented show that the recombinant cetacean PVs contain a unique set of motifs in the regulatory region, indicating that upon recombination these PVs have followed a particular evolutionary path. Presumably, these motifs evolved as an adaptive response to the need of additional/modified regulation for effective gene expression/replication/packaging of these chimeric genomes. Nonetheless, in one of the recombinant PVs, TtPV2, we did not identify any of these specific motifs. TtPV2 is indeed basal to all other recombinant cetacean PVs in the early gene tree (Fig. 1), suggesting that the

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

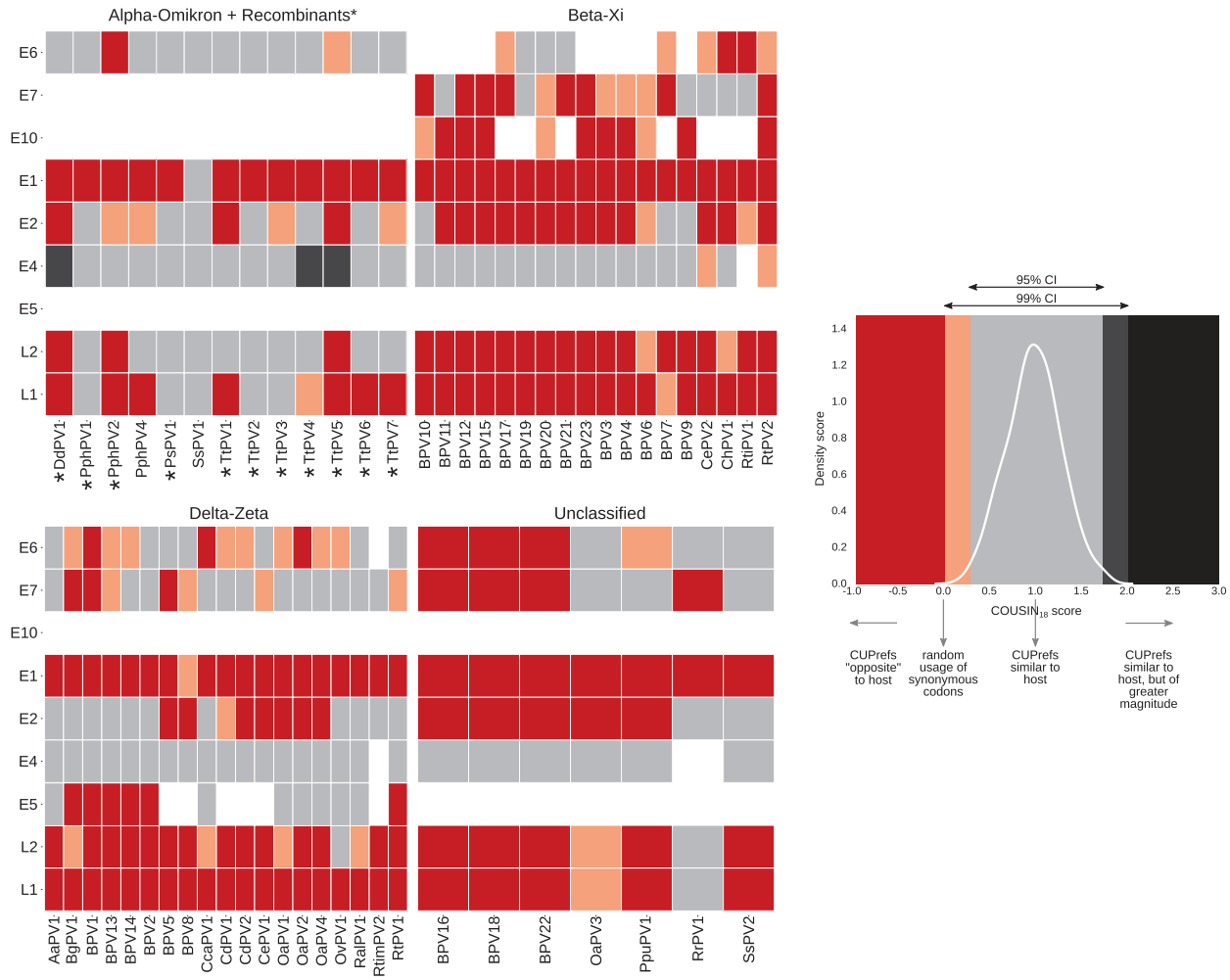


Figure 5. Heatmap of the COUSIN scores for all PV ORFs of fifty-eight PVs infecting Cetartiodactyla. COUSIN scores are stratified by PV ORFs (rows: *E6*, *E7*, *E10*—when present, *E1*, *E2*, *E4*, *E5*—when present, *L2*, and *L1*), listed in the order they are present in the PV genome, and by PV type (columns) that are grouped based on the PV crown groups: Alpha-OmikronPVs (including recombinant PVs), Beta-XiPVs, Delta-ZetaPVs, and unclassified PVs. Recombinant cetacean PVs are indicated with an asterisk. The COUSIN scores reflect the similarity between the CUPrefs in a given case gene (the corresponding viral gene) and those in a reference gene set (the full gene set in the corresponding host genome). Interpretation of the COUSIN score is given in the inset, and illustrated by colours that have been used as guideline for the heatmap. The curve in the inset corresponds to the COUSIN scores of a simulation of 500 random sequences composed of 100 codons, generated with the same CUPrefs as the different Cetartiodactyla hosts. The 95 per cent and 99 per cent confidence intervals (CIs) were calculated, and subsequently compared to the COUSIN score of the different PV ORFs. If the COUSIN score of a PV gene falls outside these intervals (coloured red, salmon, dark-grey, or black), it is considered significantly different from the reference, and when the score falls within the interval of the 95 per cent CI (coloured light grey), it is judged as matching the reference. Most viral genes display CUPrefs that are significantly different from those of the host, being systematically enriched in codons that are underrepresented in the host's genes.

appearance of the specific motifs in the URR occurred after the recombination event, as well as after the branching of TtPV2 from all other recombinant PVs. This observation supports our hypothesis of an adaptive response in the PV genome to drastic changes in the virus–host interactions associated to the recombination event.

When comparing the distribution of motifs with the evolutionary distances, we observe a better correspondence between motif composition and early genes phylogeny than with late genes phylogeny. We interpret that this agreement between early genes and motif repertoire reflects the fact that motifs in the URR are mostly involved in early gene expression regulation and genome replication, while control elements for late gene expression regulation are not located within the URR. In Alpha-Omikron PVs, the best characterized PVs, promoters for late gene expression are located within *E7* (Hummel, Hudson, and Laimins 1992; Ozbun and Meyers 1998; Bernard 2013). It is therefore not unexpected to observe such a correlation. On the contrary, it is surprising that the distribution of motifs in the URR correlates almost equally well with viral taxonomy and with host taxonomy. This suggests that besides the precise viral gene assembly, adaptation to the host species also play an important role in the evolution of regulatory PV motifs.

As the genomes of the recombinant cetacean PVs are composed of gene cassettes stemming from two distantly related viral lineages (Alpha-OmikronPVs and Beta-XiPVs), infecting distant hosts (Phocoenidae/Suidae and Bovidae/Cervidae), one could also expect to observe trends in extant gene CUPrefs, so that orthologous genes from viruses infecting closely related hosts would display closer CUPrefs than those infecting distantly related hosts. Our results do not show particular differences in CUPrefs between recombinant and non-recombinant PVs

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

Table 3. Comparison of the CUPrefs, motif distribution, and phylogenetic clustering with viral taxonomy, host taxonomy, sampling location, and clinical presentation.

	ORF or genomic region	Viral taxonomy (four categories)		Host taxonomy (seven categories)		Anatomical sampling location (five categories)		Clinical presentation (four categories)	
		Correlation	*P*-value	Correlation	*P*-value	Correlation	*P*-value	Correlation	*P*-value
Codon usage preferences	*E6*	0.1621	<0.001	0.2140	<0.001	0.1418	<0.001	0.0600	0.621
	E7	0.0700	0.005	0.1176	0.026	0.0938	0.420	0.0572	0.840
	E10	Only present in Beta-Xi	NA	0.1177	0.286	0.1982	0.500	0.2179	0.273
	E1	0.2822	<0.001	0.3528	<0.001	0.2354	<0.001	0.0635	0.206
	E2	0.2024	<0.001	0.2358	<0.001	0.1278	<0.001	0.0427	0.832
	E4	0.1610	<0.001	0.1726	<0.001	0.1594	<0.001	0.0630	0.231
	E5	Only present in Delta-Zeta	NA	0.2917	<0.001	Only one location	NA	Only one clinical pres.	NA
	L1	0.1428	<0.001	0.2060	<0.001	0.0941	0.054	0.0711	0.069
	L2	0.1419	<0.001	0.1648	0.007	0.1172	0.014	0.0732	0.085
Motif distribution	URR	0.3280	<0.001	0.2605	<0.001	0.1829	<0.001	0.0371	0.878
Phylogenetic clustering	*E1-E2*	0.5966	<0.001	0.3362	<0.001	0.2441	<0.001	0.0849	0.020
	L2-L1	0.5189	<0.001	0.2897	<0.001	0.1939	<0.001	0.0794	0.038

A PERMANOVA test was performed to test significance beyond null expectation for the respective correlation between qualitative traits (viral taxonomy: Alpha-Omikron, Beta-Xi, Delta-Zeta, and unclassified; host taxonomy: Bovidae, Camelidae, Cervidae, Delphinidae, Giraffidae, Phocoenidae, Suidae; anatomical sampling location: alimentary tract, anogenital, eye, hair follicles, and skin; clinical presentation: asymptomatic infection, benign (fibro)epithelial lesion, malignant lesion, and fluid running from eyes), and the Euclidian distances of CUPrefs of each PV ORF, the Jaccard distances of the presence/absence matrix of conserved motifs detected in the URR, and pairwise phylogenetic distances of the *E1-E2* and *L2-L1* trees. The good match between phylogenetic clustering and viral taxonomy is expected, as PV taxonomy boundaries are designed based on phylogenetic relatedness. The repertoire of motifs in the URR is more closely related to the viral taxonomy than to the host taxonomy. On the contrary, for all genes CUPrefs are better correlated with host taxonomy than with viral taxonomy.

infecting Certartiodactyla. Only for all genes in SsPV1, infecting pigs, the CUPrefs differ from those of all other PVs infecting cetartiodactyles. Otherwise, we observe that orthologous genes of PVs belonging to different crown groups display closer CUPrefs, than non-orthologous genes from the same virus, so that early and late genes tend to respectively display similar CUPrefs, independently of the viral genome. Such differences in CUPrefs between early and late genes have already been described for PVs infecting humans (Félez-Sánchez et al. 2015). Concordantly, CUPrefs in late genes are likely related to the cellular context in which they are expressed—differentiating epithelial cells—which provides with a particular tRNA pool for translation (Zhou et al. 1999).

As viruses depend on the host machinery for translation, we also assessed whether the CUPrefs of the Certartiodactyla PVs match those of the hosts they infect. As already shown for PVs infecting humans (Félez-Sánchez et al. 2015), we observe that CUPrefs of the Certartiodactyla PVs do not match those of the hosts they infect, to the extent that viral genes are systematically enriched in codons that are rare in the host's genome, and this independent from their nature of early or late genes. Overall, the lack of match between PVs and host CUPrefs has been explained as a strategy to avoid overexposure to the immune system (Tindle 2002). Only in certain PVs, the *E4* gene displays CUPrefs closer to those of the host, whereas for *E5* the CUPrefs appear linked to those of the hosts. While little is known about the expression pattern of E5, the E4 protein is usually expressed at high levels, and interacts with cytoskeletal proteins facilitating virion release (Doorbar 2013). The differences in CUPrefs between PV ORFs relative to the CUPrefs of their hosts, suggests that also for Certartiodactyla PVs CUPrefs are linked to gene expression patterns as well as gene function, as proposed for human PVs (Félez-Sánchez et al. 2015).

SsPV1 is the only PV that clearly distinguishes itself from other Certartiodactyle PVs in terms of CUPrefs (Figs 4 and 5). This virus was isolated from different individual stabled pigs (Stevens et al. 2008), and has also been detected in pig slurry (Di Bonito et al. 2019). Differences in CUPrefs between SsPV1 and SsPV2 can be related to the presentation of the infection, as SsPV1 has been isolated from healthy skin (Stevens et al. 2008), while SsPV2 has been isolated from papillomatous lesions in wild boars (Link et al. 2017). This result matches again previous observations on human PVs showing that differences in CUPrefs correspond well to the different clinical presentations (Félez-Sánchez et al. 2015).

In summary, we have shown here that recombination in PVs infecting Certartiodactyla occurred most probably through one single recombination event. This event generated 'chimeric' genomes of distantly related PVs. As an adaptive response to this drastic change in genome composition and in cellular context for gene expression, new regulatory motifs evolved in the URR of recombinant PV genomes. A gene expression study among cetacean PVs could shed light on the adaptive phenotypes that were affected by the changes in regulatory motifs observed in this study.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

Acknowledgements

We are grateful to the *genotoul* bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources. The authors acknowledge the IRD itrop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this article.

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

Funding

This work was supported by the European Research Council Consolidator Grant CODOVIREVOL (contract number 647916 to I.G.B.) and by the European Union Horizon 2020 Marie Sklodowska-Curie research and innovation programme grant ONCOGENEVOL (contract number 750180 to A.W.).

References

Angulo, M., and Carvajal-Rodríguez, A. (2007) 'Evidence of Recombination within Human Alpha-Papillomavirus', *Virology Journal*, 4: 33.

Antonsson, A., and Hansson, B. G. (2002) 'Healthy Skin of Many Animal Species Harbors Papillomaviruses Which Are Closely Related to Their Human Counterparts', *Journal of Virology*, 76: 12537–42.

Bahir, I. et al (2009) 'Viral Adaptation to Host: A Proteome-Based Analysis of Codon Usage and Amino Acid Preferences', *Molecular Systems Biology*, 5: 311.

Bailey, T. L et al. (2009) 'MEME SUITE: tools for Motif Discovery and Searching', *Nucleic Acids Research*, 37: W202–8.

Bennett, M. D. et al (2008) 'Genomic Characterization of a Novel Virus Found in Papillomatous Lesions from a Southern Brown Bandicoot (*Isoodon obesulus*) in Western Australia', *Virology*, 376: 173–82.

Bergvall, M., Melendy, T., and Archambault, J. (2013) 'The E1 Proteins', *Virology*, 445: 35–56.

Bernard, H.-U. (2013) 'Regulatory Elements in the Viral Genome', *Virology*, 445: 197–204.

Di Bonito, P. et al (2019) 'Evidence for Swine and Human Papillomavirus in Pig Slurry in Italy', *Journal of Applied Microbiology*, 127: 1246–54.

Bourret, J., Alizon, S., and Bravo, I. G. (2019) 'COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences', *Genome Biology and Evolution*, 11: 3523–8.

Bravo, I. G., and Alonso, A. (2004) 'Mucosal Human Papillomaviruses Encode Four Different E5 Proteins Whose Chemistry and Phylogeny Correlate with Malignant or Benign Growth', *Journal of Virology*, 78: 13613–26.

Bravo, I. G., and Felez-Sanchez, M. (2015) 'Papillomaviruses: Viral Evolution, Cancer and Evolutionary Medicine', *Evolution, Medicine, and Public Health*, 2015: 32–51.

Castresana, J. (2000) 'Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis', *Molecular Biology and Evolution*, 17: 540–52.

Doorbar, J. (2013) 'The E4 Protein; Structure, Function and Patterns of Expression', *Virology*, 445: 80–98.

Van Doorslaer, K. et al (2018) 'ICTV Virus Taxonomy Profile: Papillomaviridae', *Journal of General Virology*, 99: 989–90.

Félez-Sánchez, M. et al (2015) 'Cancer, Warts, or Asymptomatic Infections: Clinical Presentation Matches Codon Usage Preferences in Human Papillomaviruses', *Genome Biology and Evolution*, 7: 2117–35.

Gottschling, M. et al (2011a) 'Modular Organizations of Novel Cetacean Papillomaviruses', *Molecular Phylogenetics and Evolution*, 59: 34–42.

Gottschling, M. et al (2011b) 'Quantifying the Phylodynamic Forces Driving Papillomavirus Evolution', *Molecular Biology and Evolution*, 28: 2101–13.

Gupta, S. et al (2007) 'Quantifying Similarity between Motifs', *Genome Biology*, 8: R24.

Hummel, M., Hudson, J. B., and Laimins, L. A. (1992) 'Differentiation-Induced and Constitutive Transcription of Human Papillomavirus Type 31b in Cell Lines Containing Viral Episomes', *Journal of Virology*, 66: 6070–80.

Johansson, C. et al (2012) 'HPV-16 E2 Contributes to Induction of HPV-16 Late Gene Expression by Inhibiting Early Polyadenylation', *The EMBO Journal*, 31: 3212–27.

Link, E. K. et al (2017) '*Sus scrofa* Papillomavirus 2-Genetic Characterization of a Novel Suid Papillomavirus from Wild Boar in Germany', *Journal of General Virology*, 98: 2113–7.

López-Bueno, A. et al (2016) 'Concurrence of Iridovirus, Polyomavirus and a Unique Member of a New Group of Fish Papillomaviruses in Lymphocystis Disease Affected Gilthead Seabream', *Journal of Virology*, 90: 8768–79.

McBride, A. A. (2013) 'The Papillomavirus E2 Proteins', *Virology*, 445: 57–79.

Narechania, A. et al (2005) 'Phylogenetic Incongruence among Oncogenic Genital Alpha Human Papillomaviruses', *Journal of Virology*, 79: 15503–10.

Ozbun, M. A., and Meyers, C. (1998) 'Temporal Usage of Multiple Promoters during the Life Cycle of Human Papillomavirus Type 31b', *Journal of Virology*, 72: 2715–22.

R Core Team (2018) 'R: A language and environment for statistical computing', R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rector, A. et al (2008) 'Genomic Characterization of Novel Dolphin Papillomaviruses Provides Indications for Recombination within the Papillomaviridae', *Virology*, 378: 151–61.

Rector, A., and Van Ranst, M. (2013) 'Animal Papillomaviruses', *Virology*, 445: 213–23.

Robinson, D. F., and Foulds, L. R. (1981) 'Comparison of Phylogenetic Trees', Mathematical Biosciences, 53: 131–47.

Robles-Sikisaka, R. et al (2012) 'Evidence of Recombination and Positive Selection in Cetacean Papillomaviruses', *Virology*, 427: 189–97.

Soria-Carrasco, V. et al (2007) 'The K Tree Score: Quantification of Differences in the Relative Branch Length and Topology of Phylogenetic Trees', *Bioinformatics*, 23: 2954–6.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.

Stevens, H. et al (2008) 'Isolation and Cloning of Two Variant Papillomaviruses from Domestic Pigs: *Sus scrofa* Papillomaviruses Type 1 Variants a and b', *Journal of General Virology*, 89: 2475–81.

Tindle, R. W. (2002) 'Immune Evasion in Human Papillomavirus-Associated Cervical Cancer', *Nature Reviews Cancer*, 2: 59–64.

Willemsen, A., and Bravo, I. G. (2020) 'Ecological Opportunity as a Driving Force of Radiation Events and Time-Dependent Evolutionary Rates in Papillomaviruses,' bioRxiv. 2020.03.08.982421. doi: 10.1101/2020.03.08.982421.

Wingender, E. (2000) 'TRANSFAC: An Integrated System for Gene Expression regulation', *Nucleic Acids Research*, 28: 316–9.

Woolford, L. et al (2007) 'A Novel Virus Detected in Papillomas and Carcinomas of the Endangered Western Barred Bandicoot (*Perameles bougainville*) Exhibits Genomic Features of Both the Papillomaviridae and Polyomaviridae', *Journal of Virology*, 81: 13280–90.

Zhou, J. et al (1999) 'Papillomavirus Capsid Protein Expression Level Depends on the Match between Codon Usage and tRNA Availability', *Journal of Virology*, 73: 4972–82.

Downloaded from https://academic.oup.com/ve/article/6/1/veaa038/5831844 by guest on 02 September 2021

Chapter 3

Subfunctionalisation of Paralogous Genes and Evolution of Differential Codon Usage Preferences : The Showcase Of Polypyrimidine Tract Binding Proteins

The third chapter of this manuscript is a result of collaboration between a fellow doctorate student (now acquired PhD), Jérome Bourret, our supervisor Ignacio G. Bravo and myself. It is accessible on BiorXiv as a preprint, and is to be submitted for peer review in the immediate future.

In this chapter we propose that over time gene paralogs can evolve to have divergent CUPrefs, that allow for their differential expression in space and time. We use the example of PTBPs – the Polypyrimidin tract binding proteins, encoded by a number of genes found in all vertebrates. These genes are present as three main paralogs, PTBP1, PTBP2 and PTBP3. It was shown by Robinson and collaborators, that in humans these paralogs have different expression patterns linked to their different CUPrefs(Figure 1)(Robinson et al., 2008). This differential evolution of gene expression timing could be interpreted as the result of selective pressures being released on the duplicated gene as the original can still maintain its function, while the duplicate can explore new functions or expression patterns.

Here we study the PTBP paralogs of 47 mammalian and 27 non-mammalian Vertebrates as well as three protostome species as outgroups. Fifteen of these vertebrate species were more deeply studied as we could retrieve well-annotated full genomes. We use clustering methods, Phylogenetic reconstruction, ancestral state reconstruction, and looked at forces that may have shaped the CUPrefs of these genes.

We observed that PTBP1s are the GC richest paralogs, while PTBP3s are the AT-richest ones. Moreover, there is a significant difference in GC content between PTBP1s of mammalian *vs* non-mammalian vertebrates, mammalian species being enriched in GC. This is further confirmed by k-means and hierarchical clustering : when analyzing the full CUPrefs, we retrieve three groups : PTBP1 in mammals, PTBP1 in non-mammals, and all PTBP2 & PTBP3 clustering together.

In the 15 well-annotated species, we inspected the genomic context in order to assess whether the observed GC richness in mammals is a result of local mutational forces. Therefore we compared GC3 content of the paralogs to their flanking regions and introns. We found that variations in local genomic context explains almost completely the variations in GC3 content of PTBP1 (sequential least squares regression, R^2=0.97), relatively well in the case of PTBP2 (R^2=0.46), and less than half

in the case of PTBP3 (R^2=0.16). We interpret, that in the case of mammals, there is an overall GC enrichment that is clearly visible in PTBP1, but does not explain the CUPrefs of PTBP2 and PTP3. In fact the COUSIN values of mammalian PTBPs show that PTBP1s overmatch the CUPrefs of the organisms, while PTBP2 and PTBP3 undermatch it, meaning they have an increased frequency of rare codons. In their turn, non mammalian PTBPs however display CUPrefs matching to their organisms.

Phylogenetic reconstruction groups the sequences first by paralogy, then by species, hinting that two duplication events took place around the time of the emergence of vertebrates. Ancestral reconstruction and the analysis of sequences shows that mammals accumulated a large number of synonymous and non-synonymous mutations that enrich the sequences in GC, compared to non-mammalian sequences.

We explored the nature of paralogous gene evolution and their CUPrefs, and showed that PTBPs display divergent nucleotide composition and CUPrefs over the clade of vertebrates. We propose that this phenomenon is compatible with the theory of genotypic evolution by sub-functionalisation upon gene duplication. It would be interesting to explore further the expression levels of the different paralogs and its link to CUPrefs with a series of cell culture experiments, which we invite readers to do.

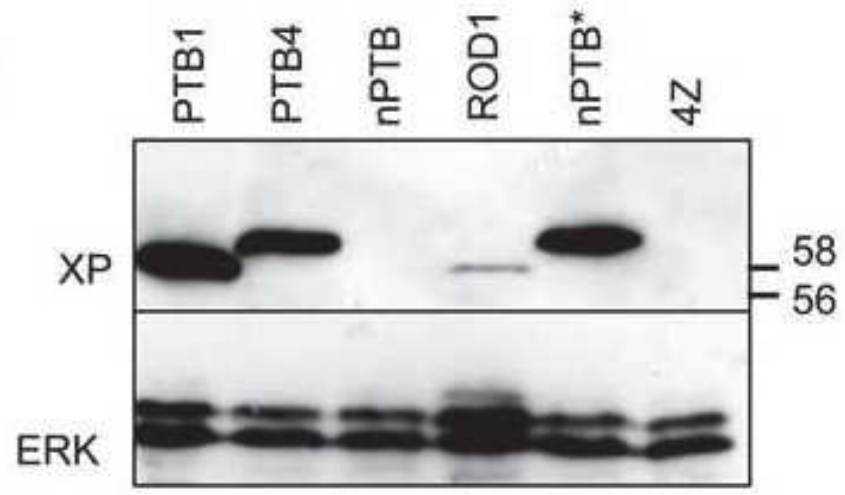


Figure 1: Differential overexpression of PTB, nPTB and ROD1 *-Western blot of HeLa cells transfected with different PTBP versions (PTBP1, PTBP4 – an isoform of PTBP1, nPTBP- alternative name for PTBP2, ROD1- alternative name for PTBP3, nPTBP* - GC enriched PTBP2), probed with anti-Xpress (XP, upper panel) or anti-ERK antibodies , from (Robinson, Jackson, & Smith, 2008).*

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

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

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

[†]These authors contributed equally to this work

ABSTRACT

Gene paralogs are copies of an ancestral gene that appear after gene or full genome duplication. When the two sister gene copies are maintained in the genome, redundancy may release certain evolutionary pressures, allowing one of them to access novel gene functions. Here we focused on the evolutionary history of the three polypyrimidine tract binding protein (*PTBP*) paralogs and their concurrent evolution of differential codon usage preferences in vertebrate species.

PTBP1-3 show high identity at the amino acid level (up to 80%), but display strongly different nucleotide composition, divergent CUPrefs and distinct tissue-specific expression levels. Phylogenetic inference suggests that the duplication events leading to the three extant *PTBP1-3* lineages predate the basal diversification within vertebrates. We identify a distinct substitution pattern towards GC3-enriching mutations in *PTBP1*, concurrent 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 directional mutation–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* are linked to rare CUPrefs and specialized spatio-temporal expression. This scenario 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 : fanni.borveto@ird.fr

1 Significance Statement

In vertebrates, PTBP paralogs display strong differences in gene composition, gene expression regulation, and their expression in cell culture depends on their codon usage preferences. We show that placental mammals PTBP1 have become GC-rich because of local mutational pressures, resulting in an enrichment of frequently used codons and in a strong, tissue-wide expression. On the contrary, PTBP2 in vertebrates are AT-rich, with a lower contribution of local mutational processes to their specific nucleotide composition, show high frequency of rare codons and in placental mammals display a restricted expression pattern contrasting to that of PTBP1. The systematic study of composition and expression patterns of gene paralogs can help understand the complex mutation-selection interplay that shape codon usage bias in multicellular organisms.

2 Introduction

During gene 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 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 with the same amino acid. These 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 prokaryotes and some 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 Vertebrates 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 upon 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 GC3-rich 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 in culture 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 in 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.

3 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) taking the human *PTBP* paralogs as references (see supplementary Material S2 for accession numbers). We could identify the corresponding three orthologs in all Vertebrate 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 from three protostome genomes, which contained a single *PTBP* homolog per genome (Supplementary Material S2). 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 synteny and composition information on the flanking regions and introns (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 the selected 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. To reduce information dimension for the analysis of 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) (Supplementary Material)

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.

4 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*s 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

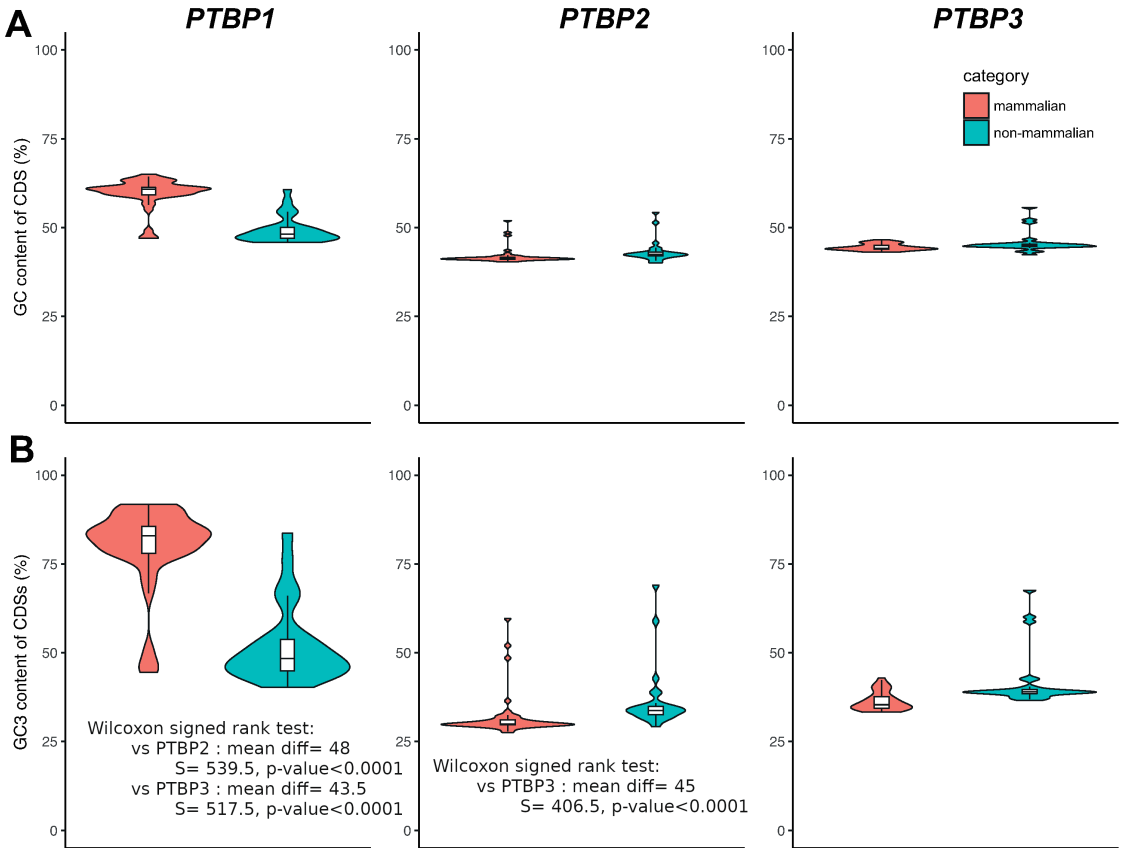


Figure 1: **GC content (A) and GC3 content (B) of Vertebrates *PTBP*s.** Violin plots display the overall distribution while box and whiskers display median, quartiles and 95% of the corresponding values for mammalian (red) and non-mammalian (blue) individual genomes. Results of Wilcoxon signed rank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

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 are 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 and are 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 of *PTBP1* (R^2=0.97) and relatively well in the case of *PTBP2* (R^2=0.46), the fraction of variance explained by the local composition significantly drops for *PTBP3* (R^2=0.15).

Vertebrate *PTBP* paralogs differ in CUPrefs

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

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 the third axes (respectively 6.7% and 3.2%). 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 protostomates *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; Kappa-Fleiss consistency score = 0.76) 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. 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-

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

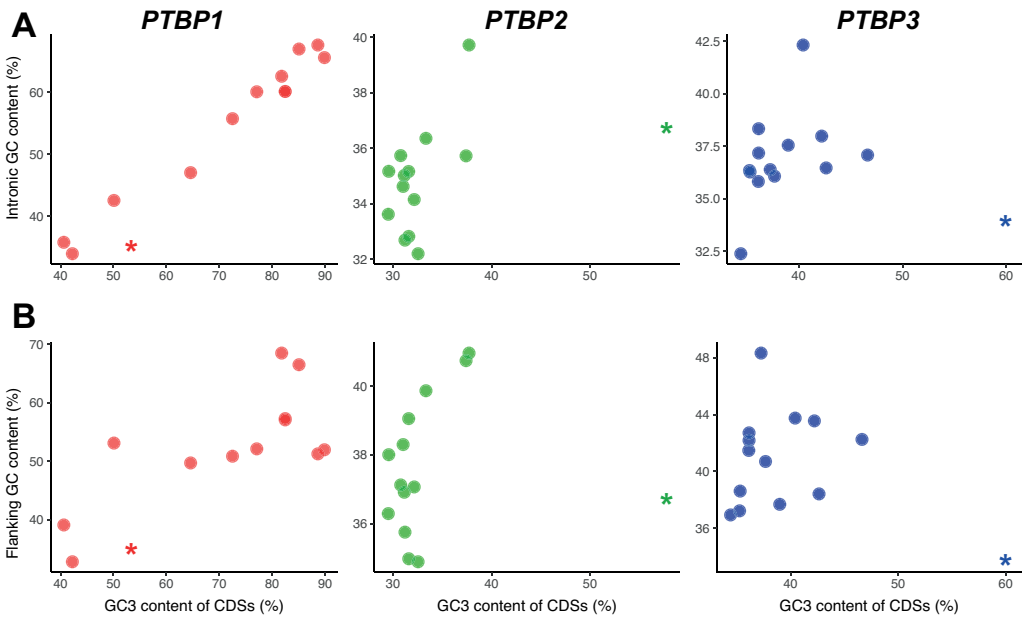


Figure 2: **Variation in GC3 content of *PTBP*s (x-axis) and in the GC content of the corresponding introns (A, y axis) or flanking regions (B, y axis)**. Each dot represents one of the 15 individual used for the genomic context analysis. The asterisk indicates the values for the species *Danio rerio*,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

ending 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 than the average in the genome, while *PTBP2*s are enriched in rare codons to the extent that their CUPrefs go in the opposite direction to the average in the genome. As for *PTBP3*, in mammals we observe COUSIN values below 0 in most cases or very close to 0 in the case of the horse *Equus caballus* and house mouse *Mus*

musculus, implying a tendency towards rare codons. In non-mammals however, PTBPs show an overall similarity to their respective reference CUPrefs.

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 (Figure 4, Supplementary Material Figure S10). 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

Table 3: Results for an individual (left) or for a sequential (right) least squares regression for explaining variation in GC3 composition of *PTBP*s genes, by variation of different local (introns or flanking regions of the corresponding gene) or of global (all coding CDS, all introns and all flanking regions in the corresponding genome) compositional variables in 14 well-annotated vertebrate genomes. For the sequential fit, variables are ordered according to their contribution to the sequentially better model, and the order may thus differ between paralogs. 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 contributions			**Sequential contribution**		
Parameter	**R^2**	**P value F test**	**Parameter**	**R^2**	**BIC**
Local_GC_intron	0.9726	<0.001	Local_GC_intron	0.9726	66.4765
Local_GC_flanking	0.5345	0.0069	Local_GC_flanking	0.974 (n.s.)	68.3142
Global_GC3_exome	0.7279	0.0004	Global_GC3_exome	0.9749 (n.s.)	70.3842
Global_GC_introns	0.116	0.2786	Global_GC_flanking	0.9803(n.s.)	69.9886
Global_GC_flanking	0.1041	0.3065	Global_GC_introns	0.9806(n.s.)	72.2531
PTBP2					
Individual contributions			**Sequential contribution**		
Parameter	**R^2**	**P value F test**	**Parameter**	**R^2**	**BIC**
Local_GC_intron	0.3738	0.0264	Local_GC_intron	0.4558	60.1257
Local_GC_flanking	0.4558	0.0113	Global_GC_introns	0.4895(n.s.)	61.8583
Global_GC3_exome	0.0943	0.3075	Global_GC3_exome	0.4914(n.s.)	64.3761
Global_GC_introns	0.0488	0.4684	Global_GC_flanking	0.4934(n.s.)	66.8894
Global_GC_flanking	0.0287	0.5801	Local_GC_flanking	0.4974(n.s.)	69.35
PTBP3					
Individual contributions			**Sequential contribution**		
Parameter	**R^2**	**P value F test**	**Parameter**	**R^2**	**BIC**
Local_GC_intron	0.1554	0.1825	Local_GC_intron	0.1554	74.7338
Local_GC_flanking	0.0522	0.4528	Local_GC_flanking	0.2095(n.s.)	76.4388
Global_GC3_exome	0.0504	0.461	Global_GC_introns	0.2718(n.s.)	77.9368
Global_GC_introns	0.0002	0.9661	Global_GC3_exome	0.2938(n.s.)	80.1032
Global_GC_flanking	0.0024	0.8744	Global_GC_flanking	0.2938(n.s.)	82.667

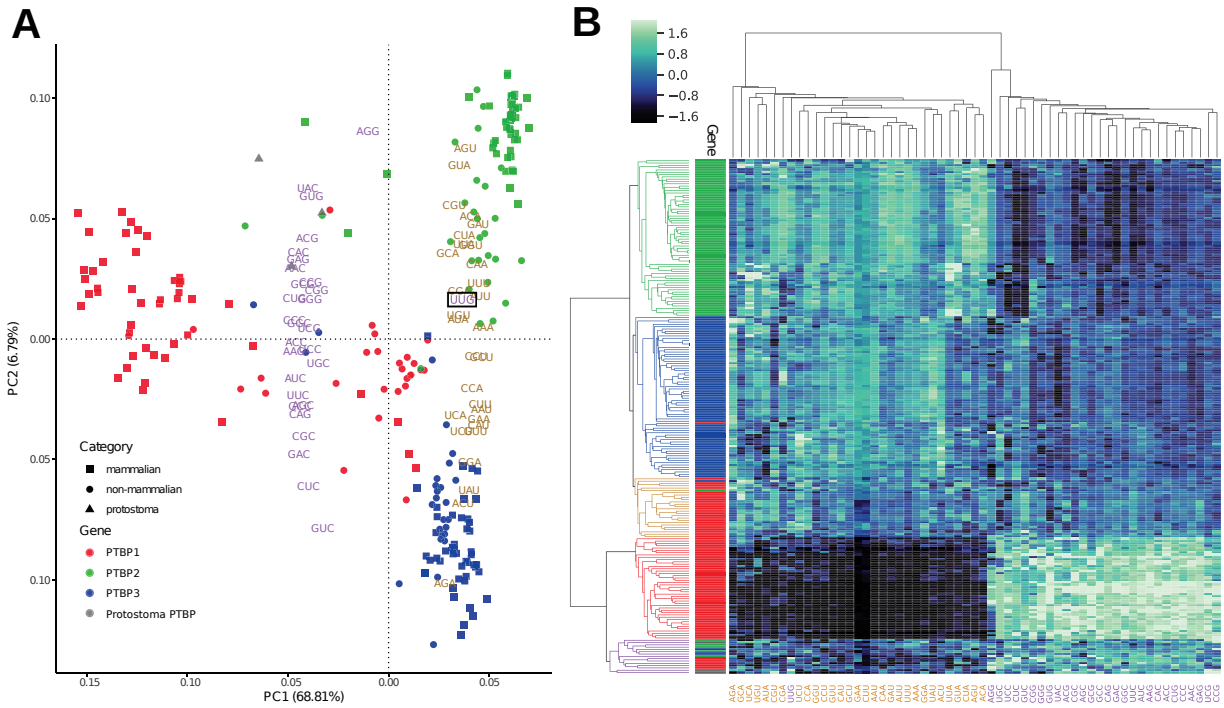


Figure 3: **CUPrefs analysis of *PTBP*s**. A) Plot of the two first dimensions of a PCA analysis based on the codon usage preferences of *PTBP1*s (red), *PTBP2*s (green), *PTBP3*s (blue) and protostoma (grey) individuals. Taxonomic information is included as mammals (squares), non-mammals (circles) and protostomes (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)

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 a very long branch, hampering the proper identification of the Vertebrate *PTBP* tree root. Amino acid and nucleotide subtrees were 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 (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*.

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 correspond to the number of partitions that are not shared between two 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

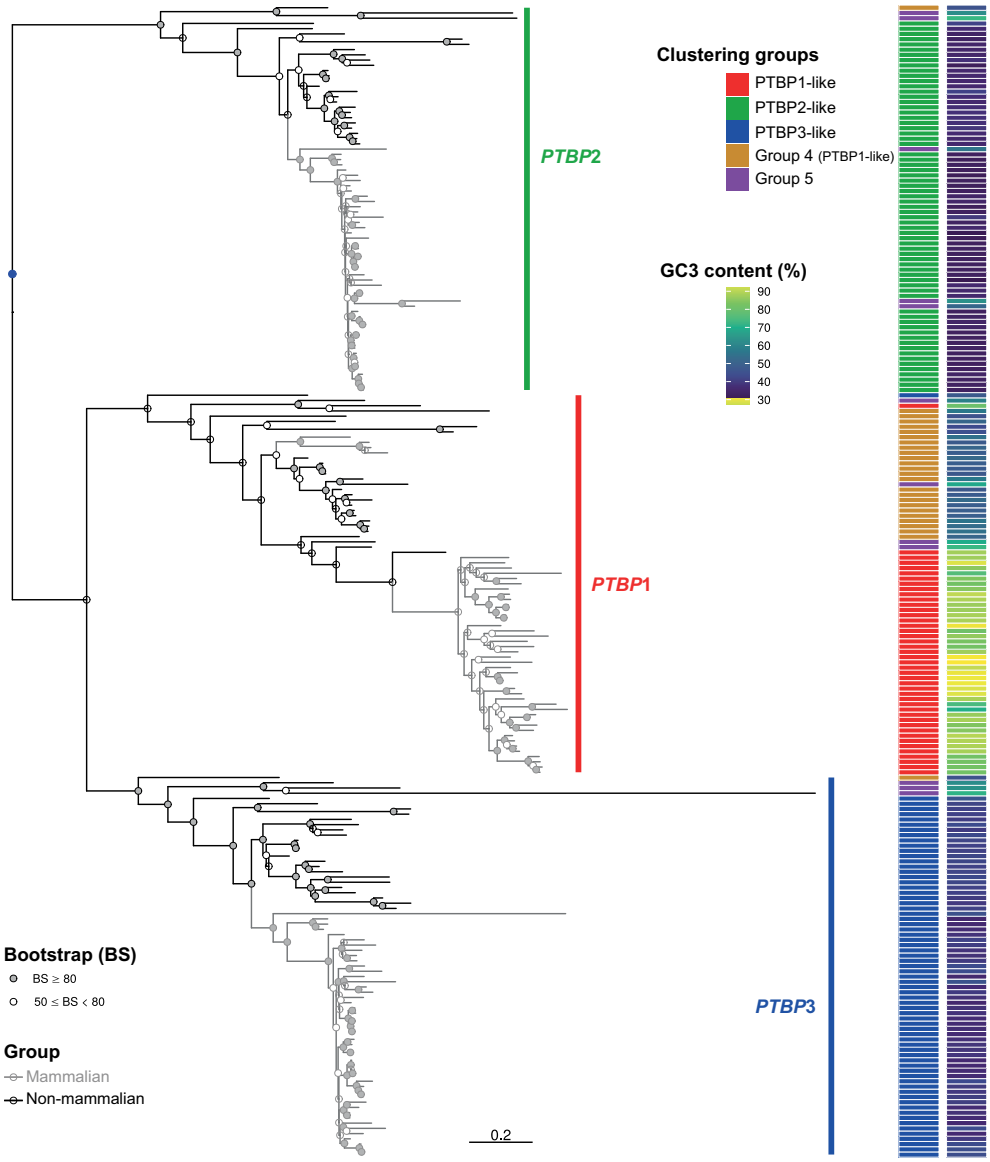


Figure 4: **Maximum-likelihood nucleic acid phylogeny of *PTBP*s genes.** The phylogram depicts *PTBP2*s (green side bar), *PTBP1*s (red side bar) and *PTBP3*s (blue side bar) clades. The outgroup genes from protostomata are not shown to focus on the scale for vertebrate *PTBP*s, but their placement 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).

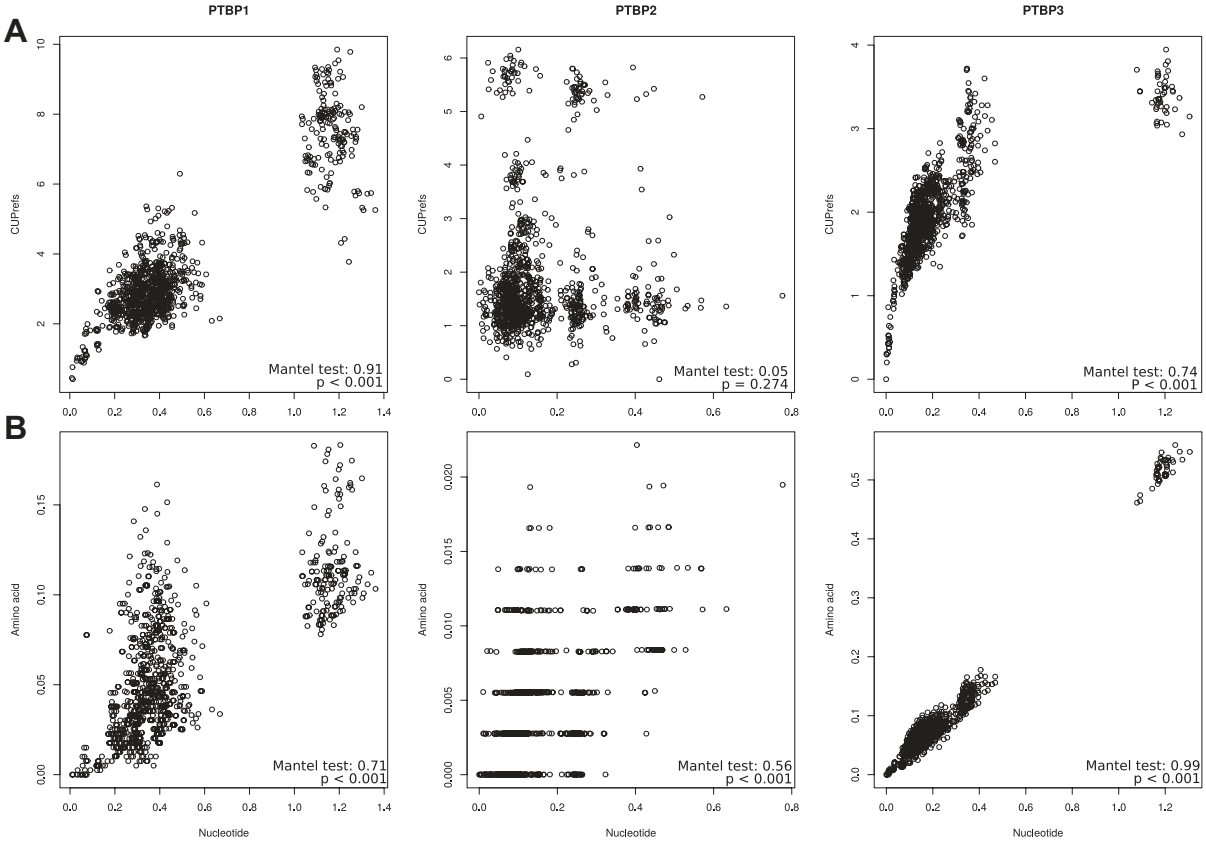


Figure 5: **Nucleotide-based pairwise distances against A) CUPrefs-based and B) amino acid-based pairwise distances for the different mammalian *PTBP* orthologs.** The results for a Mantel test assessing the correlation between the corresponding matrices are shown in each inset.

We have then analysed the correspondence between nucleotide-based and amino acid-based pairwise distances to measure the extent of codon usage bias impact on the obtained phylogeny. We observe a good correlation between both reconstructions for all paralogs, except for mammalian *PTBP2*s, which display extremely low divergence at the amino acid level (Figure 5 B, Supplementary Material S8 B). For mammalian *PTBP1*s, the plot allows to clearly differentiate a cloud with the values corresponding to the monotremes+marsupial mammals, split apart from placental mammals in terms of both amino acid and nucleotide distances. This distribution matches well the fact that monotremes+marsupials cluster separately from placental mammals in *PTBP1* phylogeny (see grey branches being paraphyletic 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* paralog, recapitulated synonymous and non-synonymous mutations between extant sequences and these ancestors, and constructed the corresponding mutation matrices (table S11). 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.

5 Discussion

The non equal use of synonymous codons has puzzled biologists since first described. It gave rise to fruitful (and unfruitful) controversies between defenders of *all-is-neutralism* and defenders of *all-is-selectionism*, and launched further the quest for additional molecular signaling beyond codons themselves (Callens et al., 2021). 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, and/or protein folding and stability. In the present work we have built on the experimental results of Robinson and coworkers, which display the differential ex-

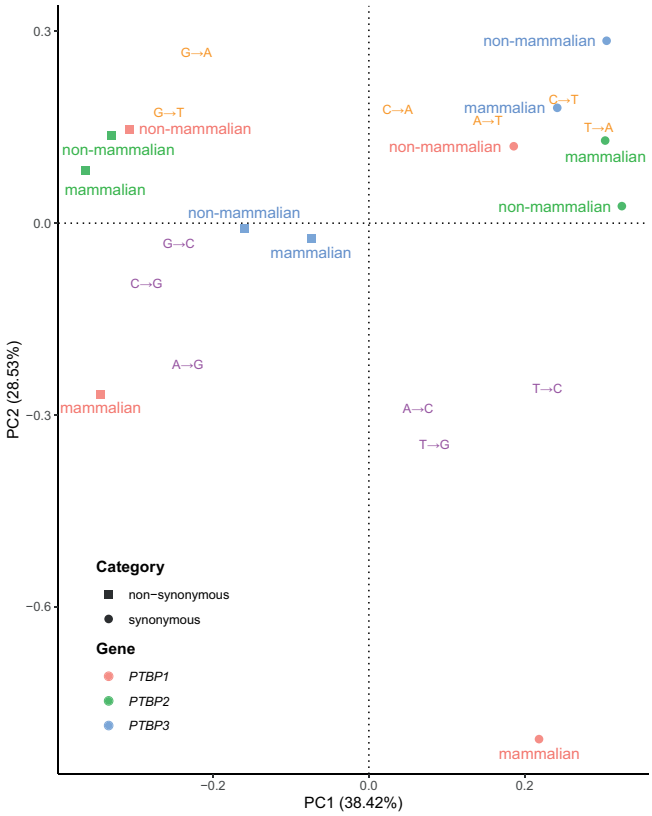


Figure 6: **Mutational spectra of synonymous and non-synonymous substitutions for *PTBP*s.** This principal component analysis (PCA) has been built using the observed nucleotide synonymous and non-synonymous substitution matrices for each *PTBP* paralog, inferred after phylogenetic inference and comparison 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).

pression 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 the nature of the connection between paralogous gene evolution and CUPrefs. Our results show that the three *PTBP* paralogous genes of Vertebrates, with divergent expression patterns, also have divergent nucleotide composition and CUPrefs. We carry on Robinson and coworkers suppositions and propose here that this evolutionary pattern could be compatible with a phenomenon of phenotypic evolution by sub-functionalisation (in this case specialisation in tissue-specific expression levels), linked to genotypic evolution by association to specific CUPrefs patterns. Such conclusions invite to pursue Robinson and coworkers efforts by comparing *PTBP*s CUPrefs-modulated expression among numerous Vertebrates cell lines, especially between mammalians and non-mammalians ones.

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 ENSEMBL and 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 protostome outgroup precludes the inference of a clear polarity between vertebrate *PTBP*s. We identify no 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*. In fact, monotremes and marsupials constitute a monophyletic clade, without placental mammals and not basal to them. Multiple findings in our results show sharp, contrasting patterns between *PTBP1* and the *PTBP2-3* paralogs: 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 as shown by the strong correlation between coding and non-coding GC content in *PTBP1* orthologs (Figure 2; Table 3).

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, overall AT3-richer without any clear correlation between coding and non-coding GC content among studied species (Figure 2; 3). 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 and GC bias gene conversion mechanism, underlying such heterogeneity, predominantly shape local nucleotide composition in numerous Vertebrates genomes, so that the physical location of a gene along the chromosome largely explains its CUPrefs (Holmquist, 1989). In agreement with this mutational bias hypothesis, variation in GC3 composition of *PTBP1*s is almost totally (R2=0.97) explained by the variation in local GC composition (Figure 2; 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 PTBP2s (R2=0.45). 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 a GC richer homolog (Marais, 2003). It has been proposed that gene expression during meiosis prevents GC-biased gene conversion during meiotic recombination (Pouyet et al., 2017). Expression of *PTBP1* in human cells is documented during meiosis in the ovocite germinal line and expression of the AT-rich *PTBP2* has been observed during spermatogenic meiosis (Zagore et al., 2015; Hannigan et al., 2017). With the assumption that *PTBP*s patterns of expression are shared between mammalian species, the GC-richness of *PTBP1* is not due to any GC-biased conversion and the low GC content of *PTBP2* can be explained by an accumulation of GC->AT and AT->AT mutations. The low GC-content observed in *PTBP3*, coupled with a lack of correlation with

neither coding nor non-coding GC-content could indicate that other mechanisms may shape the observed CUPrefs for this paralog.

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, mammalian GC3-rich *PTBP1*s match better the average genomic CUPrefs than AT3-richer *PTBP2* and *PTBP3*, which 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, consistent with an enrichment in preferentially-used codons, while on the contrary, the COUSIN value of -0.477 for *PTBP2* and of -0.235 for *PTBP3* points towards a strong enrichment in rare codons (Supplementary Material S4). The poor match between human *PTBP2* CUPrefs and the human average CUPrefs could result in low expression of these genes in different human and murine cell lines, otherwise capable of expressing *PTBP1* at high levels and *PTBP3* at a lesser degree (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 a standard practice for gene expression engineering, despite our lack of understanding the whole impact of local and global gene composition, nucleotide CUPrefs or mRNA structure on gene expression (Brule and Grayhack, 2017).

The poor expressibility of *PTBP2* in human cells, the increase in protein production by the introduction of common codons, along with mutational biases failing to explain entirely *PTBP2* nucleotide composition and CUPrefs, raise the question of the adaptive value of poor CUPrefs in 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 down-regulates *PTBP1* expression, which in turn leads to up-regulation 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 addition to local genomic context analyses, we explored *PTBP* chromosomal location and local synteny (Supplementary Material S13). The results show that, while it is clear that the position of human *PTBP1* is telomeric, and thus in one of the GC-richer region of human chromosome 9, most *PTBP*s are randomly positioned among species. Thus, while the specific location of human *PTBP1* may have influenced its CUPrefs, it is unclear whether the chro-

mosomic location of *PTBPs* have an impact on observed nucleotide composition. Synteny of *PTBP*s genes seems to be conserved, with some exceptions: most mammalian *PTBP1*s have a conserved synteny block that differs from non-mammalian species ; with the exception of *D. rerio*. *PTBP2* and *PTBP3* synteny seems conserved between mammalian and non-mammalian species with only *D. rerio* lackig the *SUSD1* gene between *PTBP3* and *UGCG*. Such results could indicate that vertebrate radiation has been followed up by a changing of *PTBP1* genomic context, with a swapping in flanking genes in mammalian branches. Such results could be related to the observed GC-content drift in *PTBP1* between mammalian and non-mammalian species.

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)), in coronavirus genomes (Daron and Bravo, 2021), 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 data-set, 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 showcase 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 potentially 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.

6 Acknowledgments

J.B. was 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 (albeit meagre) intramural support. The computational results presented have been achieved in part using the IRD Bioinformatic Cluster itrop.

7 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.

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

Callens M, Pradier L, Finnegan M, Rose C, Bedhomme S. 2021. Read between the lines: Diversity of nontranslational selection pressures on local codon usage. Genome Biology and Evolution. 13.

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

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

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

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

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

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

Daron J, Bravo IG. 2021. Variability in codon usage in coronaviruses is mainly driven by mutational bias and selective constraints on cpg dinucleotide. Viruses. 13:1800.

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

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

Duret L, Mouchiroud D. 1999, April. 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. Publisher: National Academy of Sciences Section: Biological Sciences.

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

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

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

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. 2018, May. 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.

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

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

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

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

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

Ikemura T. 1981, September. 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.

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

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

Keppetipola N, Sharma S, Li Q, Black DL. 2012, August. 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.

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

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

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

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

Laurin-Lemay S, Rodrigue N, Lartillot N, Philippe H. 2018. 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.

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

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

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

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

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

Newman ZR, Young JM, Ingolia NT, Barton GM. 2016, March. 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.

Nirenberg MW, Matthaei JH. 1961, October. 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019, April. 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. Publisher: Oxford Academic.

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

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

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

Spencer PS, Barral JM. 2012, March. Genetic code redundancy and its influence on the encoded polypeptides. Computational and Structural Biotechnology Journal. 1.

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

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

Whittle CA, Extavour CG. 2016, September. 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. Publisher: Oxford Academic.

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

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

Chapter 4

The effects of Codon Usage Preferences on Heterologous gene expression in a long-term selection experiment

Introduction

Gene expression is a necessity for all living organisms. Its complex mechanisms can be easily summarized into two steps : transcription and translation. Transcription consists of creating a negative copy of the DNA : the mRNA, one base after another. While translation is the process of 1.) reading the codons (three consecutive bases) in the mRNA, 2.) find the tRNAs with the matching anticodons that bring the corresponding amino acids and 3.) connect the amino acids into a peptide chain that eventually folds into a protein. There are 20 amino acids that are encoded by 61 codons, meaning that the genetic code is redundant. Codons that code for the same amino acid are called synonymous codons, and it has been shown that they are not used at random(Belalov & Lukashev, 2013; Nirenberg & J. Heinrich Matthaei, 1961). Organisms, tissues, or genes, often show an increased frequency of use (usually denoted as a “preference”) for one synonymous codon over the others, and the overall trends in a gene or genome are called Codon Usage Preference (CUPref) (Payne & Alvarez-Ponce, 2019)(Plotkin, Robins, & Levine, 2004). There is an ongoing debate about the origin of CUPrefs, about the extent to which it is a result of mutational bias, translational selection and drift (Agashe, Martinez-Gomez, Drummond, & Marx, 2013; Jeacock, Faria, & Horn, 2018; Quax, Claassens, Söll, & van der Oost, 2015). There is, nevertheless, no doubt about the effects of CUPrefs on individual gene expression. Thus genes with CUPrefs matching well the available tRNAs, are translated at high levels and with a high rate, while genes with CUPrefs undermatching the tRNA pool, will be translated to low levels, more slowly and inaccurately (Torrent, Chalancon, De Groot, Wuster, & Madan Babu, 2018) .

This study is but one piece of a bigger scheme, the CODOVIREVOL project (ERC-2014- CoG-647916). In this project our team and collaborators work on an ongoing investigation on how codon usage preferences of viruses may play a role in escaping from the host immune system, and thus have an adaptive value.

For this, we need first a throughout understanding of how a host cell expresses heterologous genes that may or may not match the cell’s CUPrefs. This question started to be explored in bacteria (Amorós-Moya, Bedhomme, Hermann, & Bravo, 2010; Kane, 1995; Kaur, Kumar, & Kaur, 2018), but much less in eukaryotes, and especially in mammals. Overall, eukaryotic cells have a far more

complex protein production line than prokaryotes. Apart from the compartmentalization of cell structures and functions, we can find regulatory processes, proof-reading and correcting mechanisms that ensure the potential of modulation for gene expression. Here, we launched a long term selection experiment, using a set of synonymous gene versions to better understand the immediate and long-term effects of CUPrefs on gene expression in eukaryotes, and how each step of the process is linked. As this is a long term experiment we also investigated how the cells would eventually compensate for non-optimal match of a gene that they need for survival.

We cloned five different versions of the *shble* gene connected with a P2A peptide to an enhanced Green Fluorescent Protein gene (*egfp*), into a plasmid and transfected HEK293 cells with them, making them antibiotic resistant. They were put under three different selection treatment : Bleomycin, an antibiotic they can resist by expressing the corresponding synonymous *shble* gene; Neomycin, an antibiotic the cells can resist by the expression of the unmodified *neo_tp* gene, present in the backbone of all; and without antibiotics in the media. The *shble-egfp* complex is under the control of a by a strong *Cytomegalovirus* (CMV) promoter, ensuring a high transcription level at the launch of the experiment. Gene expression was monitored by following the levels of DNA, mRNA and protein, while the cellular phenotype was monitored by quantifying fluorescence intensity and cellular ability to grow in the presence of antibiotics.

Materials and methods

Experimental evolution setup

In order to study the effects of CUPrefs on the process of protein synthesis over time, and compensation dynamics we set up a long term selection experiment. This was repeated a year later with the exact same setup, giving us overall two replicates, each one run for seven months.

In this experiment we used modified versions of the *shble* gene (GenBank: X52869.1) (Gatignol, Durand, & Tiraby, 1988), which confers resistance to the antibiotic Bleomycin. Bleomycin acts by cleaving DNA in the M and G2 phase of the cell cycle, and thus eventually kills the cell (Sikic, 1986). The SHBLE protein is a homodimer that binds two Bleomycin molecules, thus inactivating them in a 1:1 ratio (Gatignol et al., 1988). With the help of the OPTIMIZER software (Puigbò, Guzmán, Romeu, & Garcia-Vallvé, 2007) we created several synonymous versions of this *shble* gene with varying CUPrefs and mRNA folding energy, while keeping the same amino acid sequence. After several pilot experiments, we chose five of these constructions for the final set of experiment (Figure 1, Table 1).

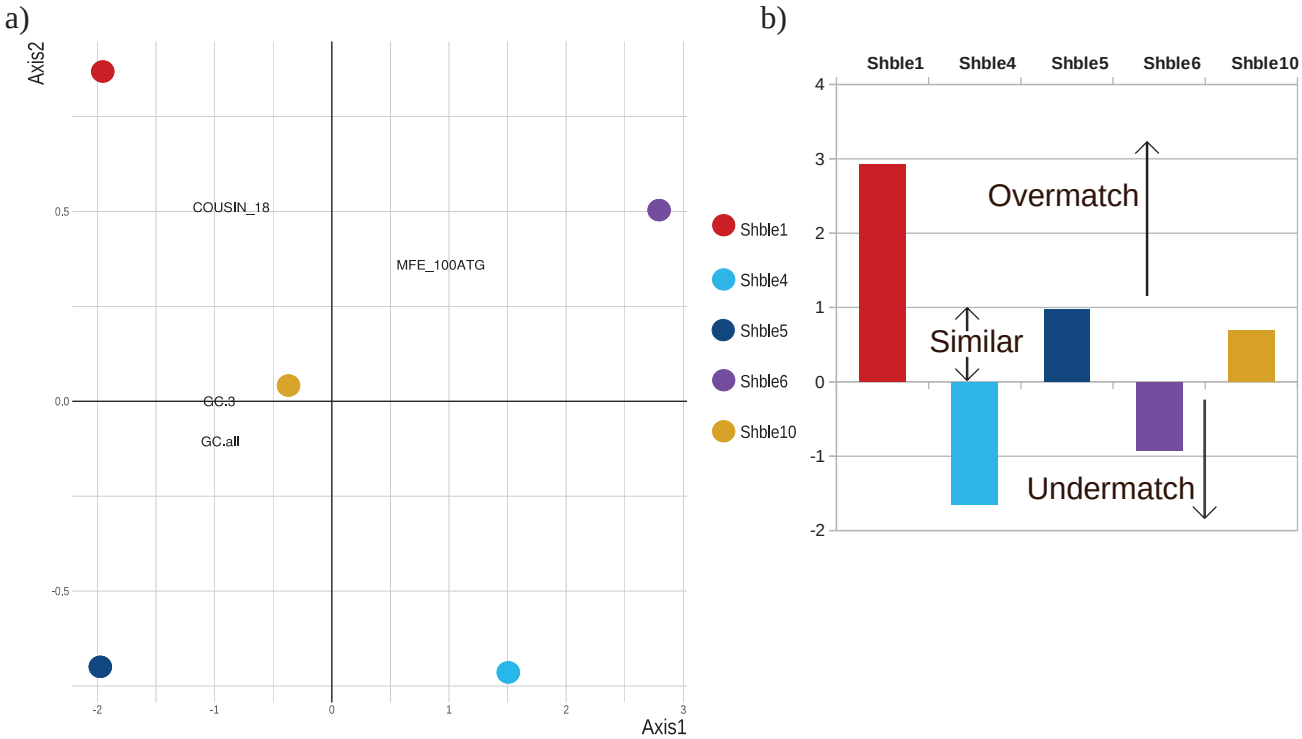


Figure 1: Comparison of created Constructs *– a) PCA of Shble1, Shble4, Shble5, Shble6 and Shble10 based on GC content, GC3, COUSIN18 score and Minimum Folding energy (MFE, +/- 50 nucleotide around ATG) . The first axis represents 89.6% of the distribution while the second axis represents 10%. Constructs are indicated with different colors. b, Histogram of COUSIN18 score in the Constructs. Shble1 has an overmatching CUPrefs, Shble5 and Shble10 have similar CUPrefs,and Shble4 and 6 have undermatching CUPrefs compared to the reference, which is the average human CUPrefs.*

Name	Description	%GC(3)	COUSIN_18
Shble1 Most Frequent	For each amino acid, the most commonly used codon in the human genome was chosen	**93.077**	**2.93**
Shble4 Least Frequent	For each amino acid, the least commonly used codon in the human genome was chosen	**33.846**	**-1.651**
Shble5 Least Frequent GC rich	For each amino acid, among the two least common codons, the one with the highest GC content was chosen	**91.538**	**0.973**
Shble6 Least Frequent AT rich	For each amino acid, among the two least common codons, the one with the lowest GC content was chosen	**9.231**	**-0.924**
Shble10 Guided Random	The overall CUPrefs are that of an average human gene, as well as the folding energy	**65.385**	**0.698**
Empty	No *shble*, only *egfp* gene and *neor*	**NA**	**NA**
Mock	WT cells without plasmid, exposed to transfection agent	**NA**	**NA**
Shble2 Most frequent GC rich	For each amino acid, among the two most common codons, the one with the highest GC content was chosen	**99.23**	**2.982**
Shble3 Most frequent AT rich	For each amino acid, among the two least common codons, the one with the lowest GC content was chosen	**20**	**-0.414**

Table 1: Characteristic of Constructs *– Description, Cousin score and GC3 content of the designed constructs. Shble1, Shble4, Shble5, Shble6 and Shble10 are the constructs used in the selection experiment along with the Mock and Empty cell lines. Shble2 and Shble3 were part of a different experiment (Day2 experiment), run before the lunch of the Selection experiment.*

Shble1-6 are one-amino-acid-one-codon forms, meaning that each and every amino acid in the corresponding *shble* sequence is always represented by the same codon. Shble10 on the other hand has been made using a guided random algorithm with the average human CUPrefs as the reference. The guided random algorithm consists of a Monte Carlo algorithm that given an amino acid, picks a codon at random based on the frequencies of use of codons in the reference (Puigbò et al., 2007). The amino acid sequence of the *shble* gene is thus back-translated using this procedure to render a coding sequence. Several shble guided-random versions were created, and classified by their high-average-low mRNA folding energy as well as by their Codon Adaptation Index scores (Sharp & Li, 1987). The mRNA minimum folding energy was calculated via the The ViennaRNA Web Service (Gruber, Bernhart, & Lorenz, 2015), taking the 50 nucleotides before and after the ATG. Shble10 was in the average for folding energy and displays a matching COUSIN score (0.698), that can be seen as a typical, anodyne human gene, as far as CUPrefs are considered. At the N-terminus of the

shble gene we introduced seven codons that provide an AU1 epitope tag after translation, allowing for Western blot detection. All *shble-egfp* complexes share thus the first eight codons, thus minimizing the differences associated to translation initiation and putting the emphasis on the effects of CUPrefs in the elongation phase.

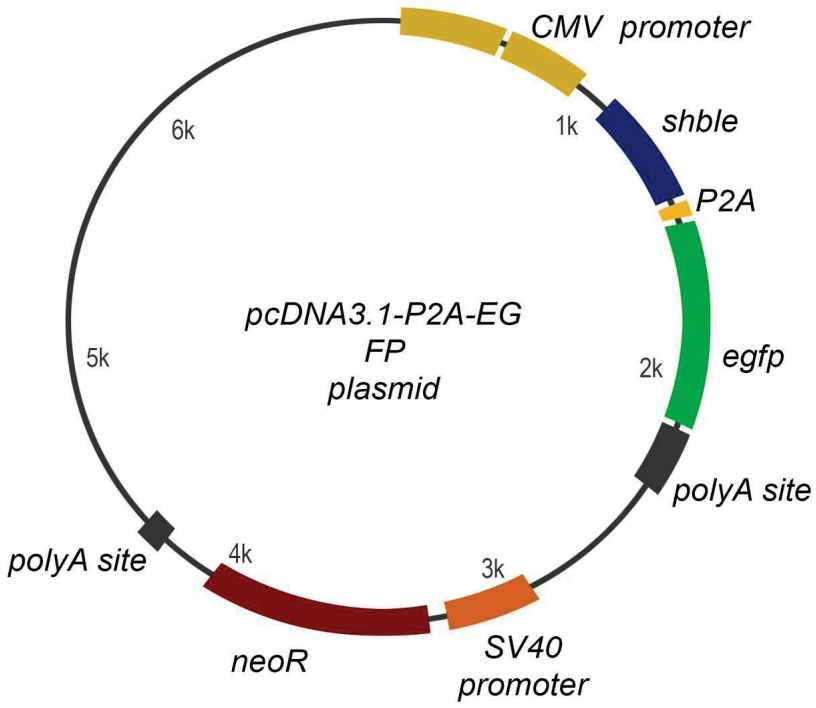


Figure 2: Details of the constructed plasmid *– Schematic representation of the modified pcDNA3.1-P2A-C-EGFP with the inserted and modified shble gene (in blue). Gene of interests, different promoters and the P2A peptide are highlighted.*

The five chosen versions cover a varying range of CUPrefs, folding energy, and COUSIN score (resemblance to a given reference, here the average CUPrefs of the human genome (Bourret et al., 2019))(Table 1). These modified *shble* sequences were then inserted in a pcDNA3.1-P2A-C-EGFP plasmid, upstream of and in frame with an *enhanced Green Fluorescent protein* (eGFP - (Cormack, Valdivia, & Falkow, 1996) gene, and flagged with AU1 epitope tag (Figure 2). The *egfp* open reading frame (ORF) is only transcribed and translated after *shble* as the *shble* and *egfp* ORF are linked with a P2A peptide. This means that at the mRNA level they form one large mRNA but at the protein level they are two functional proteins. (Figure 3) In fact the sequence coding for the P2A peptide causes the ribosome to release the nascent peptide without performing a trans-peptidation step onto the amino acid on the tRNA at the ribosome A-site, and to resume translation (Atkins et

al., 2007). Upstream the *shble-egfp* complex, is a CMV promoter, a very strong promoter originally found in the *Cytomegalovirus(Pasleau, Tocci, Leung, & Kopchick, 1985)*.

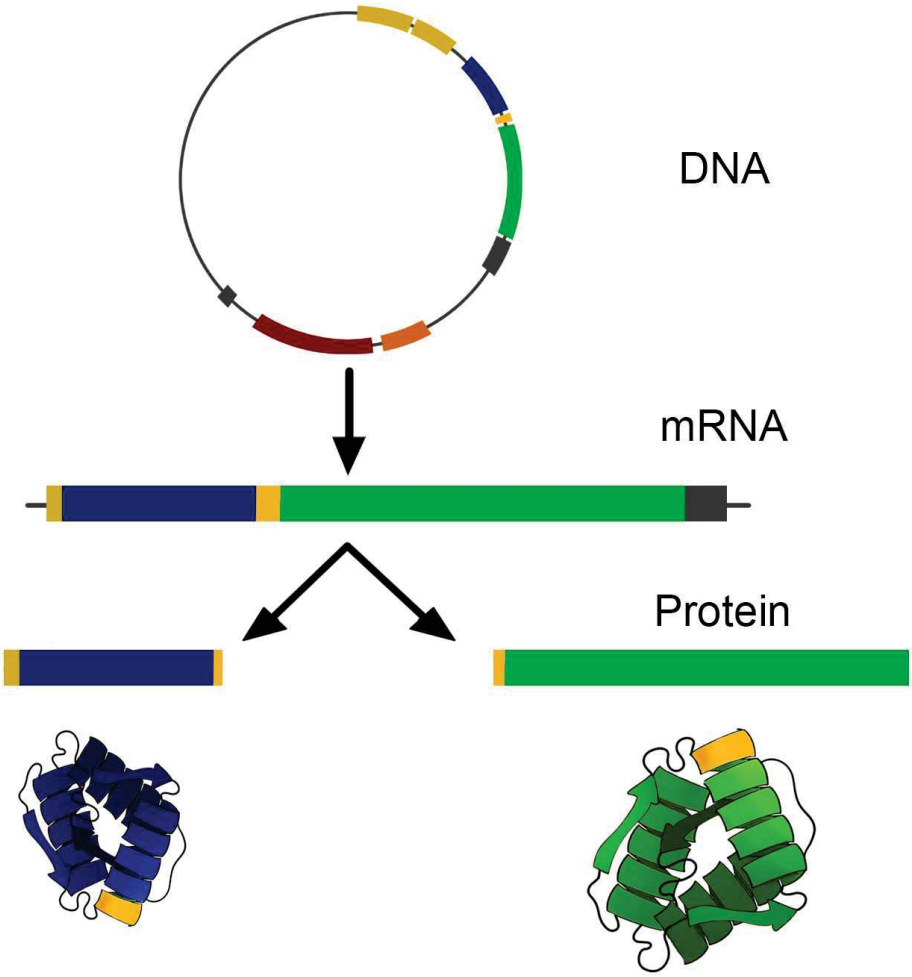


Figure 3: shble-egfp, from DNA to protein *– The shble (blue) and egfp(green) sequence is connected with a short P2A sequence (yellow) in between, on the plasmid. As a result after transcription, there is one large mRNA containing both shble and egfp. On the protein level however, because of the P2A, SHBLE and EGFP are present as two separate and functional protein, with an eventual residue of the P2A (in yellow) .*

In the backbone of this plasmid, a *neo_tp* – Neomycin resistance gene is also present under the control of an SV40 promoter (*Simian virus 40*). NEO_TP is an enzyme that inactivates Neomycin molecule by phosphorylation (Beck, Ludwig, Auerswald, Reiss, & Schaller, 1982) Overall the plasmids confer a resistance to Neomycin, and – depending on the versions- a more or less efficient resistance to Bleomycin, as well as the expression of *egfp* after every *shble*.

After analyzing the RNA Seq data we realized that Shble4 and Shble6 present a splicing site (see RNAseq results), that was not predicted by any algorithms we used (Human splicing finder (Desmet

et al., 2009), SPLM (Softberry – www.softberry.com). Shble4 presents one alternative spliced form containing an intron between positions 229 and 536. Shble6 presents two alternate forms differing in only one codon from positions 226 or 229 to 536 (Figure 4).

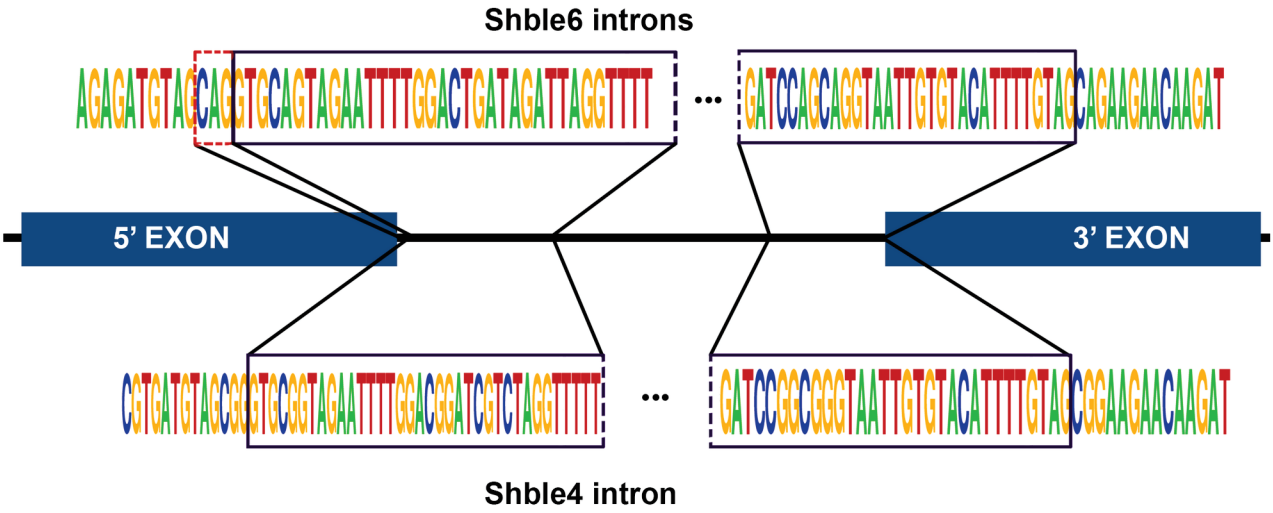


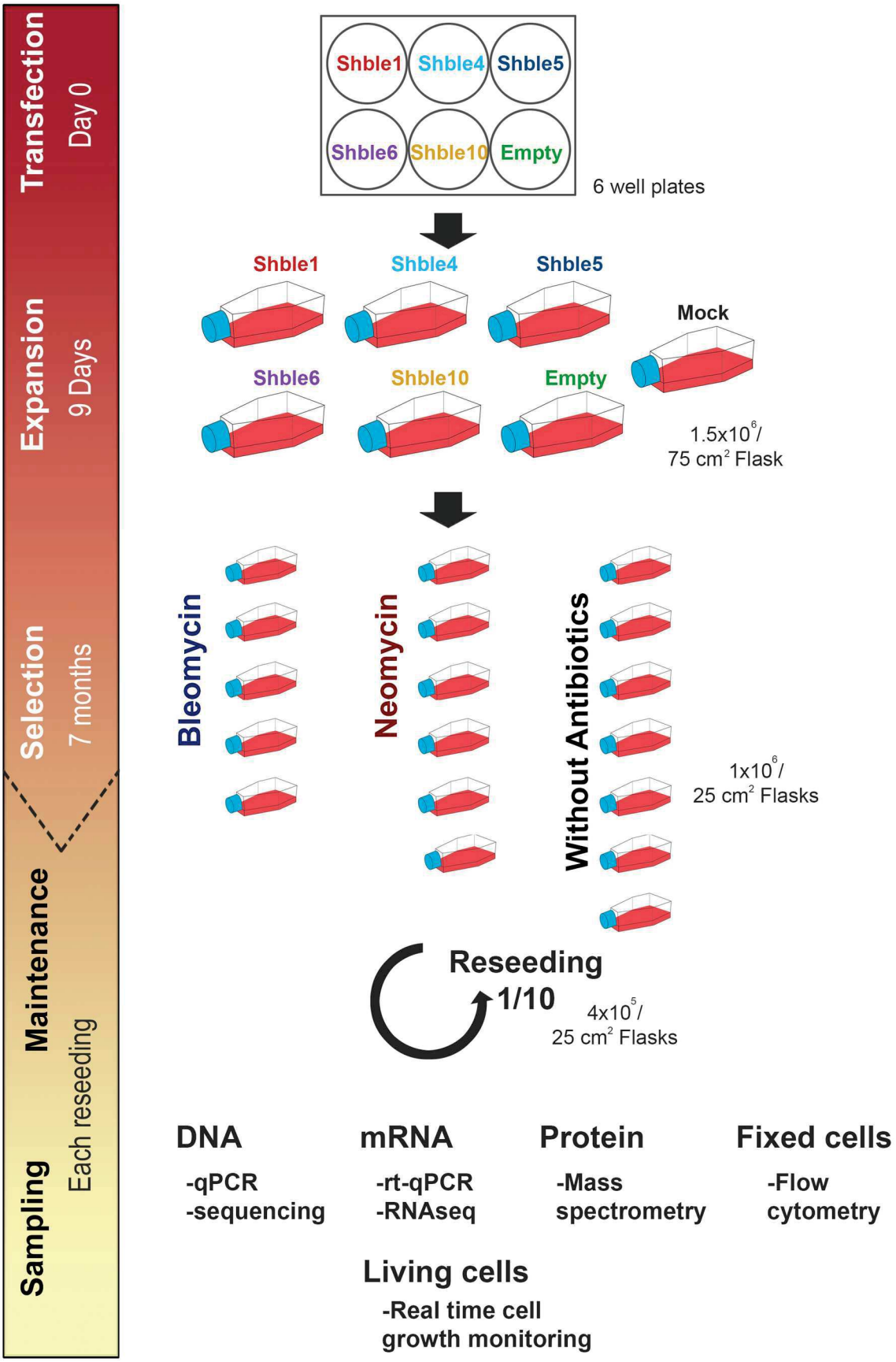
Figure 4: Splicing in Shble4 and Shble6 *– In blue a schematic representation of the shble gene, in between the exons the beginning and end of the introns (framed sequence logo) of Shble4 (below) and Shble6 (above). Shble6 has two different spliced forms, that differ in one codon at the 5' end, this is marked with a red dashed line .*

Transfection, maintenance

These five different plasmids as well as a control we called "Empty", containing the pcDNA3.1-P2A-C-EGFP plasmid without the *shble* variants, were introduced into HEK293 cells (ATCC number : CRL-1573) (Graham, Smiley, Russell, & Nairn, 1977; Thomas & Smart, 2005). The "Empty" plasmid lacks the *shble* gene, therefore, it still confers resistance to Neomycin, but not to Bleomycin. HEK293 cells are surface adherent, hypotriploid human cells. Their modal chromosome number is 64, occurring in 30% of cells with a rate of 4.2 % of cells with higher ploidies. For transfection we used the TurboFect (ThermoFisher) transfection reagent and $1.5x10^6$ HEK293 cells per construct in 6-well plates. We also created a "Mock" cell line, which was treated with the transfection agent but without any plasmid. This way all cell lines encountered the same stress at the launch. With the Empty and Mock controls, overall a total of seven cells lines were created. After letting the cells expand for nine days after transfection, we harvested them and divided each cell line into three groups of one million cells. The first group was put under selection in the presence of 400 µg/mL Bleomycin (Fisher scientific), the second in the presence of 400 µg/mL Neomycin (Fisher scientific)and finally, the third group in simple cell medium without antibiotic

selection. These cell lines were maintained for ~7 months (100-120 generations, depending on the cell line), and the experiment was completely repeated once (Figure 5).

Figure 5: Design of the Selection experiment *– The experimental setup of started with the transfection of 1x10⁶ HEK293 cells per construct in 6 well plates (4 well/ construct).Cells were then transferred to T75 flasks to increase population size for 9 days. Each population was then divided into three groups and placed in T25 flasks (1.5x10⁶ cells / flask), and put under selection. Bleomycin – with 400µg/mL of Bleomycin in the media, Neomycin – with 400µg/mL if Neomycin in the media, and a third treatment with no antibiotic selection in the cell media. Each time cells reached 90% confluence, they were reseeded at 1/10th rate and sampled for different measures as described.*


Transfection
Day 0
Expansion
9 Days
Selection
7 months
Maintenance
Each reseeding
Sampling
Shble1
Shble4
Shble5
Shble6
Shble10
Empty
6 well plates
Mock
1.5x10^6/
75 cm² Flask
Bleomycin
Neomycin
Without Antibiotics
1x10^6/
25 cm² Flasks
Reseeding
1/10
4x10^5/
25 cm² Flasks
DNA
-qPCR
-sequencing
mRNA
-rt-qPCR
-RNAseq
Protein
-Mass
spectrometry
Fixed cells
-Flow
cytometry
Living cells
-Real time cell
growth monitoring

Cells were maintained in 25mL flasks, with 5mL MEM (Minimum Essential Media – ThermoFisher) enriched with 10% FBS (Fetal Bovine Serum, Eurobio), and 1% Penicillin-Streptomycin (Fisher Scientific), at 37°C, with 5% CO2 and the corresponding antibiotic. Cells were reseeded at 1/10 dilution every time they reached 80-90% confluence and we took samples from each reseeding. The samples are named after the time-point in which they were harvested so that S0 corresponds to the moment where the cells were placed under selection, S1 is the reseeding when the cells first reached 90% confluence under selection, and so on until S30 which corresponds to the last harvest, after 30 reseeding events.

Sampling

Old medium was taken out, then the cells were rinsed with PBS solution(phosphate buffered saline, Eurobio). This buffer solution is a water-based salt solution consisting of di-sodium hydrogen phosphate and sodium chloride. As it has the same osmolarity as the human body and a pH of 7.4, it doesn't disturb the cells when used for rinsing. 1mL of trypsin (Eurobio) was added to re-suspend the adherent cells, and cells were incubated at 37°C for 5-10 minutes. Trypsine is a serine protease that cleaves the membrane proteins that adhere the cells to the surface of the flask. We added 3mL of cell media with 10% FBS to inactivate the trypsin, as too much exposition to it may damage the cells. Then we mechanically re-suspended the cells by thorough but gentle pipetting.

At the moment of reseeding the remaining $9/10^{th}$ of the population were divided into samples for further analysis. Cell pellets were taken for proteomics as well as for DNA and mRNA extraction and stored at -80 °C until treated, and $4x10^5$ cells were fixed with PBS + 2% paraformaldehyde (Fisher scientific) to be analyzed by flow cytometry,

In the second replicate we also took living cells to be used in a real-time growth measure experiment under Bleomycin. In this side-experiment, we put cells into an xCelligence machine with varying antibiotic concentration and measured their growth for 72 hours every 15 minutes (detailed below).

Sample treating

DNA extraction was performed with the Maxwell® 16 LEV Blood DNA Kit on the Maxwell 16 machine, following the manufacturer's protocol. DNA samples were later used for quantitative PCR (qPCR) and sequencing. DNA concentration was evaluated using fluorometric quantification in a Qbit machine (ThermoFisher, dsDNA HS Assay Kit).

RNA extraction was done using the Monarch Total RNA Miniprep Kit followed by a TURBO DNA-*free* (ThermoFisher)treatment as described in the protocol provided by the manufacturer. RNA extracts were later used for RNAseq and rt-qPCR. RNA concentration and RIN (RNA integrity number) were also evaluated using fluorometric quantification (ThermoFisher, Qubit RNA HS Assay Kit) and with High-Resolution Automated Electrophoresis in a Bioanalyzer system (Agilent).

For proteomics, we stored the cells in RIPA (Radioimmunoprecipitation assay buffer, composition in table S4) at -80°C until they were centrifuged overnight still in RIPA before quantification by the Bradford protein assay (Sigma-Aldrich) (Bradford, 1976). RIPA is a lysis buffer shown to enhance results for proteomics measures of nuclear, cytoplasmic and mitochondrial proteins (Ngoka, 2008). Samples were then transported to the Functional Proteomics Platform of Montpellier for unlabeled quantitative proteomics characterization.

Cells fixed for flow cytometry determination were transported to and analyzed at the MRI platform's Novocyte ACEA flow cytometer each week after sampling. Unfortunately because of the 2020 lockdown situation in France, we lack flow cytometry results for S6-S14 of Replicate two (R2) experiments. Other measurements were made without problem (qPCR, RNAseq, etc) for all time-points.

Day2 experiment

In a previous experiment performed in the team with a similar setup, HEK293 cells were transfected the same way as described before with constructs Shble1, Shble2, Shble3, Shble4, Shble5 and Shble6. Shble2 uses the most abundant GC rich codons, and Shble3 the most abundant AT rich ones. Cells were then harvested two days after transfection with the same sampling protocol. The overall procedure was repeated three times. RNAseq and proteomics results were analyzed by Marion Picard and Arthur Jallet. As some of the conclusions of this experiment are complementary to the Selection experiment I will briefly present them later.

Data preparation and analysis

Fluorescence Intensity Data

In order to analyze the flow cytometry data, we developed an R pipeline to clean and standardize the output of the cytometer machine (code to be available online). This pipeline is based on the classic, manual methods for cleaning cytometry analyses output data, and it reproduces its steps : it removes debris, outliers, and doublets (Figure 6). The advantage of this pipeline is that thresholds are adapted and applied to each file individually and automatically based on each file's distribution, in contrast to other methods, where thresholds are set based on one reference sample and applied to all. The pipeline also allows to process several hundred files at once in a few minutes (~300 ".csv" files of 3 MB in 10-20 minutes depending on CPU) and to rapidly detect samples with aberrant outputs, or unusual patterns.

When measuring individual cellular fluorescence in the FITC green channel, we also register the variables Forward scatter (FSC) and Side scatter (SSC) recording their maximum (FSC.H, SSC.H) and total value (area below the curve – SSC.A, FSC.A). Forward scatter reflects the size of the detected object (event), while side scatter reflects the internal complexity of the detected object. Here's how each step of the developed pipeline works :

<u>Removing Debris :</u>

When we remove debris, we are essentially removing values associated to events that are smaller and less complex than a standard cell would be. For this we calculate the density of the distribution of both the FSC.H (maximum Forward scatter value) and the SSC.H (maximum side scatter value) of events in a sample. In an average measure we observe a density plot with two peaks : one that is small, followed by a valley and a larger one. The second peak corresponds to the average cells, while the smaller peaks is what we call debris (mostly cellular fragments, apoptotic bodies, or dead cells). We then look for local minima, and remove all events below the local minimum before the highest peak for both FSC.H and SSC.H. Of course not all samples display this standard behavior. For the exceptions, we calculate the average threshold of debris based on the other measures from the same experiment with the same cell types and we use this value to filter out the debris. In average the fraction of debris was about 4% of the detected events.

Removing outliers :

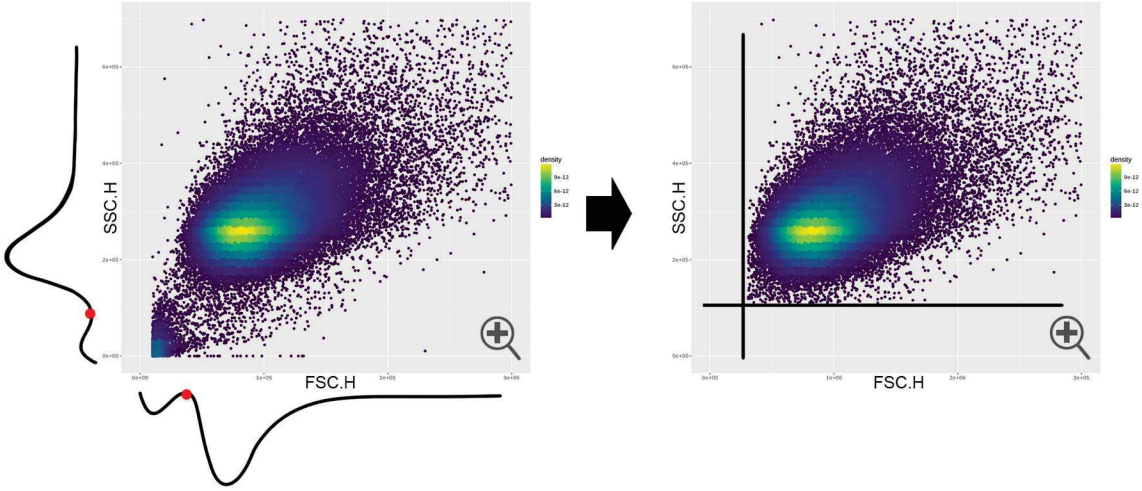
To remove outliers we use the Winsorizing method which consists in removing the events in the dataset with extreme values for the variable of interest. This is a transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers. We chose to remove 0.5% of the events with the highest values and 0.5% of the events with the lowest values of each variable (FSC.H, FSC.A, SSC.H, SSC.A), but this threshold can be easily changed in the code.

Removing Doublets:

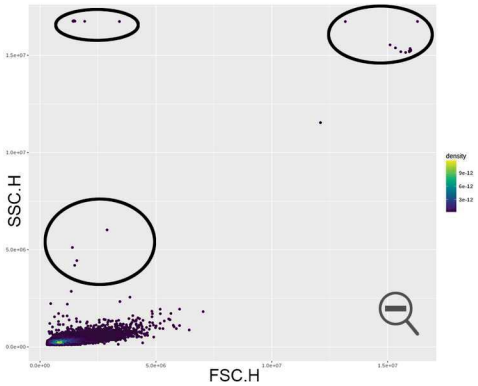
Doublets are cells that are stuck together and render distorted values in cytometry, and need thus to be removed. We can easily detect them by looking at their internal complexity (SSC), as a doublet of cells will generate in average an SSC-H signal similar to that of a single cell, but an SSC.A signal twice as big as that of a single cell. We use the same method as to remove debris, but on the SSC.H/SSC.A values. We calculate density and detect local means to determine the threshold and remove doublets from the data.

Figure 6: Flow cytometry data cleaning steps – functioning of the flow cytometry cleaning R pipeline. Color represents the density of events in the plot (blue- low, yellow – high), graphs are zoomed int for step I. and III., and zoomed out in step II. to show outliers. I. Removing Debris : Debris is removed by calculating the density of FSC.H and SSC.H. as shown in the graph with black lines. Local minima can be detected in the density plots, that separate the average cells from Debris that is smaller and less complex. Every event under the determined local minimum is removed from the data as shown. II. Removing outliers : Outliers (events circled in black) are removed by Winsorizing, we chose to remove 0.5% of the events with the highest values and 0.5% of the events with the lowest values of each variable (FSC.H, FSC.A, SSC.H, SSC.A). III. Removing doublets : Doublets are removed by calculating the density of SSC.H/SSC.A of events, and identifying the local minimum that separates doublets from average individual cells. All events below this threshold are removed from the dataset.

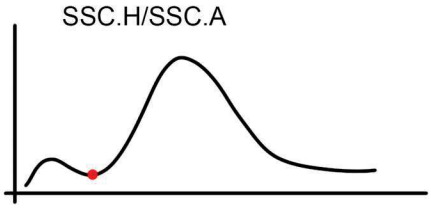
I. Removing Debris

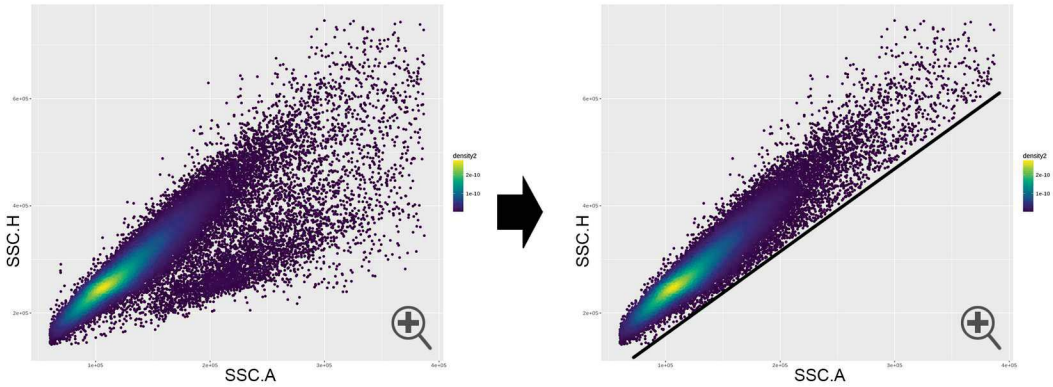


II. Removing Outliers



III. Removing Doublets





Once data were cleaned, each data point was tagged by its corresponding construct name, treatment and time-point. As this type of data are quite complex due to its dimension in time, and the amount of data for each sample, we needed to test a wide range of methods to analyze it in depth. In order to easily visualize the data and to follow the changes while the experiments were still running, we associated intervals to each cytometry event based on the green fluorescence intensity, and we plotted the varying proportions of these intervals over time for each construct in each treatment (Figure 7). This gave us a first glance into the changes that happened on the EGFP expression levels over time in our populations.

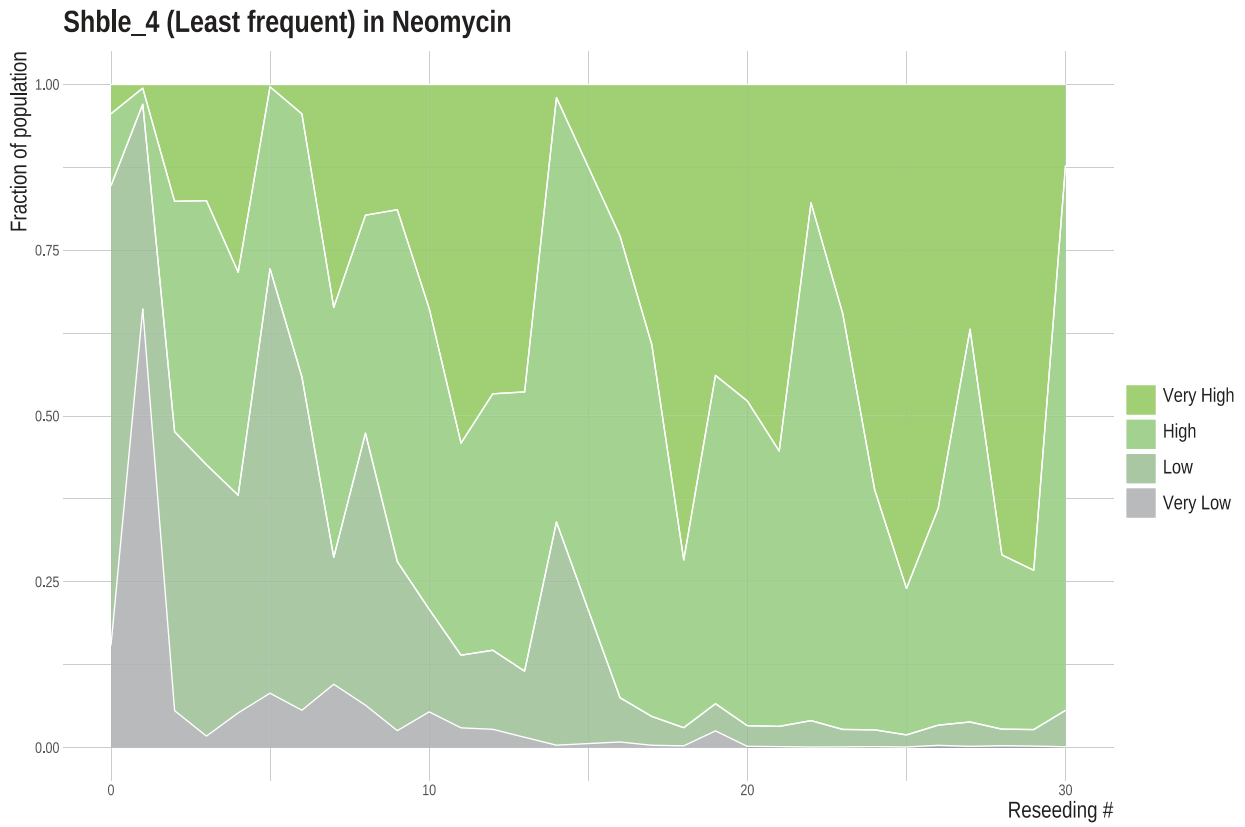


Figure 7: Example of representation of the green fluorescence intensity over time using intervals *– The fluorescence of R1 Shble 4 population under Neomycin over time (x axis : number of reseeding), represented as fraction of populations belonging to fluorescence intensity intervals (y axis): Very high > 10^6 , High > 10^5, low > 10^4 , very low < 10^4 .*

Other representations such as ridge-line graphs helped us understand that the data were even more complex than expected, as in several of our cell lines we noted the apparition of sub-populations with different fluorescence intensity levels (Figure 8). In order to explore how to best describe this multi modality of our results we calculated a number of different indicators. We first calculated the Huber M-estimator (Huber, P. J.1981) of the fluorescence of each sample, then we calculated the

medium of the 20% least and most fluorescent cells in a sample (Figure 9). We also explored how the total intensity of fluorescence in each sample represented the full population. For this we randomly sampled 19000 events from each sample, and summed up the measures of green fluorescence intensity values. We found that this total value is very highly correlated with the Huber central value (Spearman's rank correlation rho = 0.972, p-value < 2.2e-16), so we decided to keep using the Huber central value. Preliminary visualization of fluorescence data allowed us to see that there might have been a technical problem with the sampling of S30 of Replicate one (R1), so it is excluded from most analysis and we use S28 instead as the last time point of R1.

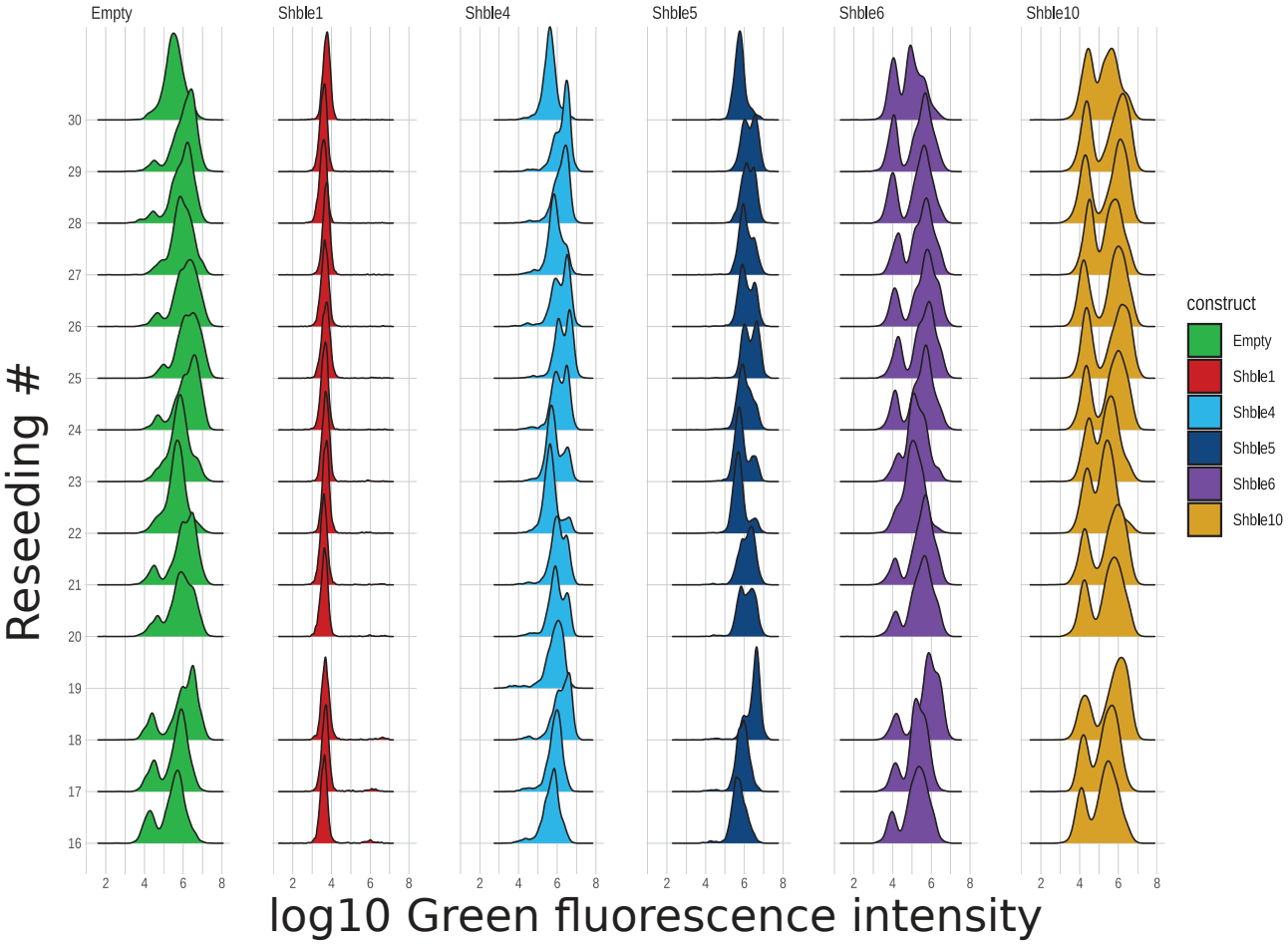


Figure 8: Example of ridge-line graph representation of green fluoresce intensity over time *– Density of population by fluorescence intensity (x axis – in log10) and by time point (Y axis) for each construct under Neomycin in R1.*

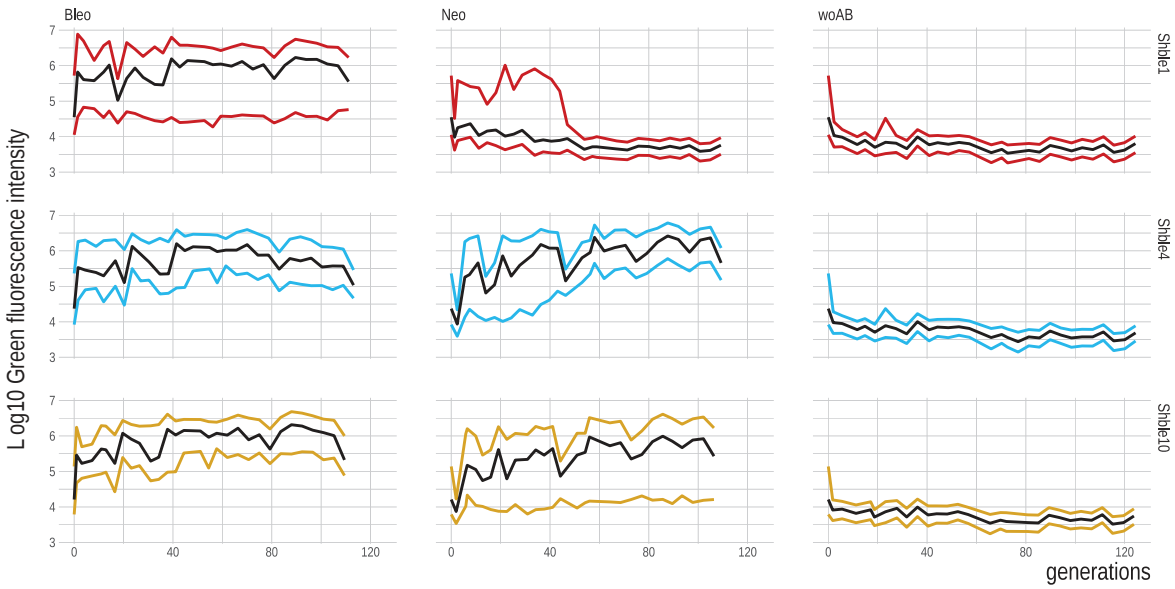


Figure 9: Example of representation of green fluoresce intensity over time by central value and upper and lower values *– green fluorescence intensity over generations of shble1, Shble4 and Shble10 under the three different treatments (Bleomycin, Neomycin, without Antibiotic). Colored lines represent the median of the 20% most (above black line) and least (below black line) fluorescent cells in the population. The Huber central value is represented with a black line. All fluorescence intensity values are in log10.*

To more easily assess differences between constructs, we carried out a series of pairwise comparisons between the beginning of the experiment and the end of the experiment. This was performed by an ANOVA followed by a *post hoc* Tukey test. As we are dealing with a very large sample size (up to 40000 data points for some samples) both the ANOVA and the *post hoc* Tukey test render very often significant differences between populations even if the size of the difference is small. For instance, we identify a significant difference in fluorescence between the mock (non-fluorescent control cells) at the beginning and the end of the experiment. As statistically significant doesn't necessary means significant from the point of view of our research, we analyzed the effect size, to quantify these differences. Cohen's D, calculated as follows (Cohen, 1988):

$$Cohen's\, d = (\bar{X}_1 - \bar{X}_2)/s$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means of group 1 and group 2, respectively, and *s* is the standard deviation of the population from which the two groups were taken, assuming a normal distribution. A d around 0.2 or less, is generally considered as "negligible" even with a significant p-value, while 0.5 indicates a "medium" effect size, and over 0.8 is "large" effect size. Other authors added the "very large" and "huge" effect size threshold, and they also remind that these thresholds are

situation-dependent, and are not carved to stone (Sawilowsky, 2009) , and that they may help us to better understand relations between groups in our data, even if in our case they stem from multimodal distributions (Table 2).

Cohen's d	effect size
0.01	*Very small*
0.2	*Small*
0.5	*Medium*
0.8	*Large*
1.2	*Very large*
2	*Huge*

Table 2: Cohen's d effect size thresholds *– based on (Sawilowsky, 2009)*

qPCR, and rt-qPCR Data

We used quantitative PCR (qPCR) to estimate the amount of targeted plasmid sequence present in our samples, and rt-qPCR to estimate the amount of targeted mRNA in the same samples. Both qPCR and rt-qPCR were run on a LightCycler® 96 real-time PCR system (*Roche*). Here we used relative quantification, meaning that we normalized all our samples with respect to a reference. We chose to analyze six time points. In the case of R1 these points were : 3, 5, 7, 13, 20 and 28 (see corresponding generations) while for R2 we have 3, 5, 7, 13 20, and 30. The reason of mismatch at the last time-point is based on observation of the fluorescence levels, in fact we suspect a technical problem with the sampling of S30 of R1.

Both qPCR and rt-qPCR were run with the same primers (Table 3, Figure 10), and the analysis was done in the same way, as the data structure doesn't differ. In each plate for each sample we ran a duplicate of three targets : a housekeeping gene –Beta tubulin-, the AU1 region -corresponding to the beginning of *shble*- and the P2A region -corresponding to the beginning of *egfp*. Additionally each plate contained standards, negative controls, and positive controls for each primer which also served as calibrators as they were the same sample in each plate. The calibrator in this case was the extracted DNA (or RNA) from a batch of HEK293 cells, transfected with the Shble1 construct, and sampled two days after transfection. We calculated the $2^{-\Delta\Delta CT}$ for each primer pair, of each sample. The Ct value (for cycle threshold), is the PCR cycle number at which the fluorescence in the sample becomes distinguishable from the background noise. Thus we normalized values first to those of the housekeeping gene in the same sample and then to those of the calibrator in the same plate.

$$\Delta Ct = Ct(gene\ of\ interest) - Ct(housekeeping\ gene)$$
$$\Delta\Delta Ct = \Delta Ct(Sample) - \Delta Ct(Calibrator)$$
$$Fold\ gene\ expression = 2^{-\Delta\Delta Ct}$$

This double normalization allows us to compare plates with one another and samples with one another, even if they weren't run at the same time. The pre-treating of the results was done in the qPCR machine's own software (LightCycler® 96 System Software), then the exported data were treated in R. Some samples didn't work (due to pipetting or mixing errors) or only one of the duplicates worked, in this case either the time-point has been removed, or the lack of duplicate is indicated.

Name	R/F	Sequence
OCODO00 8-P2A-F	F	CTGGAGACGTGGAGGAGAAC
oCODO00 7-GFP-R	R	GCTTGCCGGTGGTGCAGATG
oCODO01 0-5UTR-F	F	GAGAACCCACTGCTTACTGG
oCODO01 0-5UTR-R	R	GCCACTGTGCTGGATATCTG
tub-for	F	TCCTCCACTGGTACACAGGC
tub-rev2	R	CTCCTCTTCGGCCTCCTCAC

Table 3: Primers used for qPCR and rt-qPCR *– primers used in the Selection experiment. The first two primer pairs attach on the shble-egfp sequence as shown in figure 10., the last primer pair is for the detection of the housekeeping gene coding for Beta tubulin.*

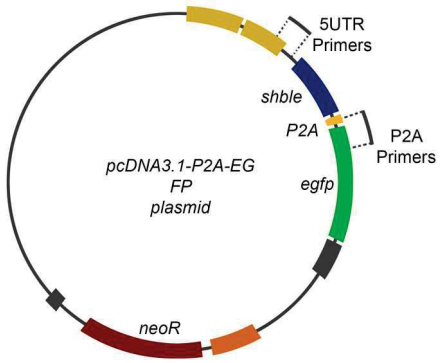


Figure 10: qPCR and rt-qPCR primers on the plasmid *– The two primer pairs used for DNA and mRNA level quantification represented as they attach on the sequence. The 5UTR primers attach just before the modified shble sequence, in the 5' UTR, while the P2A primers attach on the P2A sequence and the beginning of egfp.*

RNAseq Data

The extracted RNA samples were sent for sequencing to the company Genewiz on an Illumina Hiseq4000 instrument. Sequencing was preceded by a strand-specific RNA library preparation and a polyA selection, ensuring that the sequenced molecule were enriched in mRNAs. Sequencing was performed using paired ends at 2*150 nt. We choose to sequence four time-points for each cell line (time-point 3, 7, 13, 28) for R1 and three time-points for R2 (time-point 3, 7, 30). Once we received the raw data, Arthur Jallet and Côme Morel prepared it for analysis as described below.

First, the RNA-seq reads were quantified and aligned to the reference human transcriptome (hg38, completed with the sequences of the heterologous genes) using Kallisto (Bray, Pimentel, Melsted, & Pachter, 2016) Transcript abundances are given in transcripts per million (TPM) units. For the analysis in which we compare the transcriptome and the proteome, we collapsed all the detected mRNA isoformes of a human gene to its longest form in order to match with the proteome. Technically this means that we assigned the TPM of all possibly detected mRNA isoforms associated to a single gene to the longest form. We also excluded the non-coding RNAs that were sequenced despite the polyA selection, meaning we only took in account RNAs that can be translated to proteins, i.e. mature mRNAs. This was followed by normalization and transformations in R : we used the DESeq2 package to evaluate expression levels of protein coding genes by calculating their size factor which represents the sampling depth (Anders & Huber, 2010). This allowed us to normalize data relative to this factor and re-calculate TPM which is now normalized. This step is necessary because longer genes have higher chances to be sequenced, and this may heavily bias our results. For analyzing the evolution of the RNAseq data more over time, we used the *breseq v.0.35.5.* pipeline (Deatherage & Barrick, 2014). Breseq allows us to identify mutations in the mRNA sequences. With this method we could check if there are any mutations in the coding sequence of the heterologous genes.

Proteomics Data

In order to quantify protein molecular species, we worked with Mathilde Decourcelle and Serge Urbach from the Functional Proteomics Platform of Montpellier (CNRS). They performed label free LC-MS/MS acquisition with a nanoLC (*RSLC U3000,Thermo Fisher Scientific*) coupled to a Q Exactive HF (*Thermo Fisher Scientific*). 33 samples from R1 were analyzed this way, three time-points (S3, S13, S28) for eleven cell lines, six under Neomycin (Shble1, Shble4, Shble5, Shble6, Shble10, Empty), five under Bleomycin (Shble1, Shble4, Shble5, Shble6, Shble10). Raw results were then analyzed via the MaxQuant v1.6.10.43 and Perseus v1.6.10.43 software. These results were then transferred to us, and submitted to further cleaning, normalization and analysis.

We first removed contaminants and potential contaminants as identified by the platform, then we imputed missing values. This second step was done as there were certain proteins that were detected in some samples but not in others. This is a recurrent problem with large omic datasets, therefore missing values should normally be categorized and corrected. Missing values may be the results of errors in sample preparations, values being below the instrument’s limit of detection, or true missing

values, giving rise to three categories : Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) (Gardner & Freitas, 2020).

In our case we had 34,8% of missing values to start with, which is roughly the average in mass spectrometry-based proteomics (Gardner & Freitas, 2020). We filtered our data to keep only the proteins that were associated to at least three values (this is, proteins that were detected on at least three samples), this eliminated 250 proteins and thus reduced the percentage of missing values to 30,2%. We then identified the nature of our missing values, by comparing the intensity of proteins with missing values (ProtsMV) with that of the proteins without missing values (ProtsWoMV). We found that ProtsMVs had in general a lower intensity than ProtsWoMVs, which indicates that our missing values probably belong in the MNAR category, corresponding to proteins that are below the limit of detection. We used the QRILC method (quantile regression imputation of left-censored data) (Wei et al., 2018) to impute missing values in R via the DEP package. This method uses quantile regression to build a truncated distribution, then picks random values from it.

After correcting for the missing values we normalized the intensity-Based Absolute Quantification (iBAQ) values by the total of iBAQ values in a sample. This allows us to compare between samples, even if overall protein levels were different. This created our final variable for proteomics, the relative iBAQ (riBAQ) that we used in the follow up analyses, performed in R.

Real-time cell growth measure Data

To estimate the fitness cost of carrying the plasmid, and the associated selective advantage, we performed real-time cell growth measures in varying antibiotic concentrations. For this we used an xCelligence machine, that allows us to monitor the changes in impedance measure on gold wires on the bottom of a 96-cell culture well caused by the adherent cells as they grow and occupy more surface in the well. As the impedance is dependent on the density, size, adherence and morphology of cells, it provides a good overall measure of cell growth, that is specific of a given cell line. Cell growth was followed without antibiotic, and in presence of 400 µg/mL and 2000 µg/mL of Bleomycin. Each well contained 30,000 cells, the experiment was run for 72 hours, and repeated five times for cells coming from the Neomycin selection and six times for cell selected under Bleomycin.

Results

We started the Selection experiment after we had already acquired results from the Day2 experiment. Consequently we will first take a brief look at the Day2 RNAseq and Proteomics data analyzed following the same protocol as described above.

Two days after transfection the heterologous genes represent up to 1-7% of the whole transcriptome, while at the protein level they make up between 0.3-2.5% of the proteome.

Overall we observed that TPM values were the highest in Empty and Shble3, while Shble1, 2 and 5 have similar values, and Shble4 and especially 6 have low values (Figure 11a). Meanwhile at the protein level, SHBLE and EGFP rIBAQ values correlate, but with SHBLE values being three times lower than EGFP. The GC rich constructs (Shble1, 2, 5) showed high expression levels, while AT rich constructs displayed lower expression levels with Shble3 and 6 being close to zero (Figure 11c-d). The percentage of spliced forms present in Shble4 and Shble6 is respectively 35% and 80% two days after transfection (Figure 11b).

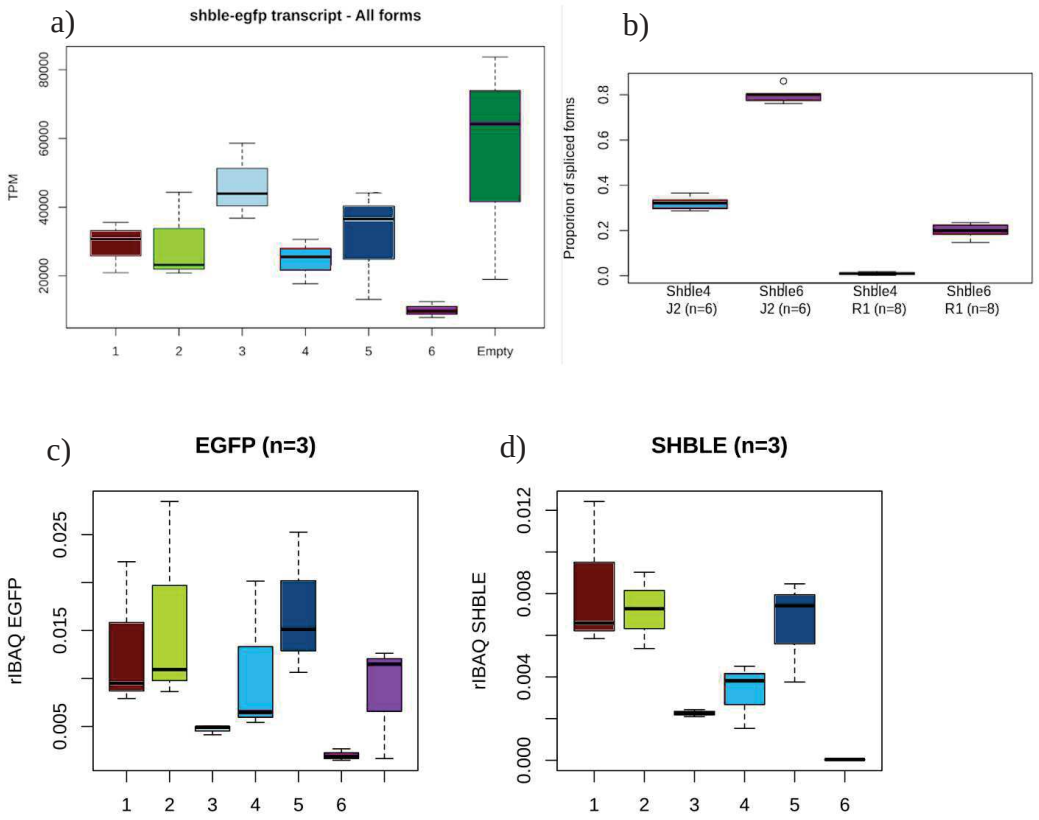


Figure 11: Results of the Day2 experiment *– a) mRNA levels (TPM) of Constructs two days after transfection b) Spliced forms of Shble4 and Shble6 two days after transfection (first two boxplot) and in the Selection experiment (last two boxplot) c) EGFP Protein levels (rIBAQ) of Constructs two days after transfection d) SHBLE Protein levels (rIBAQ) of Constructs two days after transfection*

Results for Fluorescence Data

In order to monitor the variation of protein levels in our selection experiment, we measured the individual cellular fluorescence at each sampling point (once a week approximately). We observed notable differences in fluorescence levels between treatments, and between constructs.

Without pressure to express *shble-egfp*, all cell lines lose fluorescence

Very early in the experiment, all constructs, in both replicates under the woAB treatment lost fluorescence, their fluorescence values becoming indistinguishable from the mock cell line which doesn't contain the *egfp* gene. This shows that all transfected cells stopped expressing eGFP, which may be caused by the silencing of *egfp* or a loss of plasmid (Figure 12).

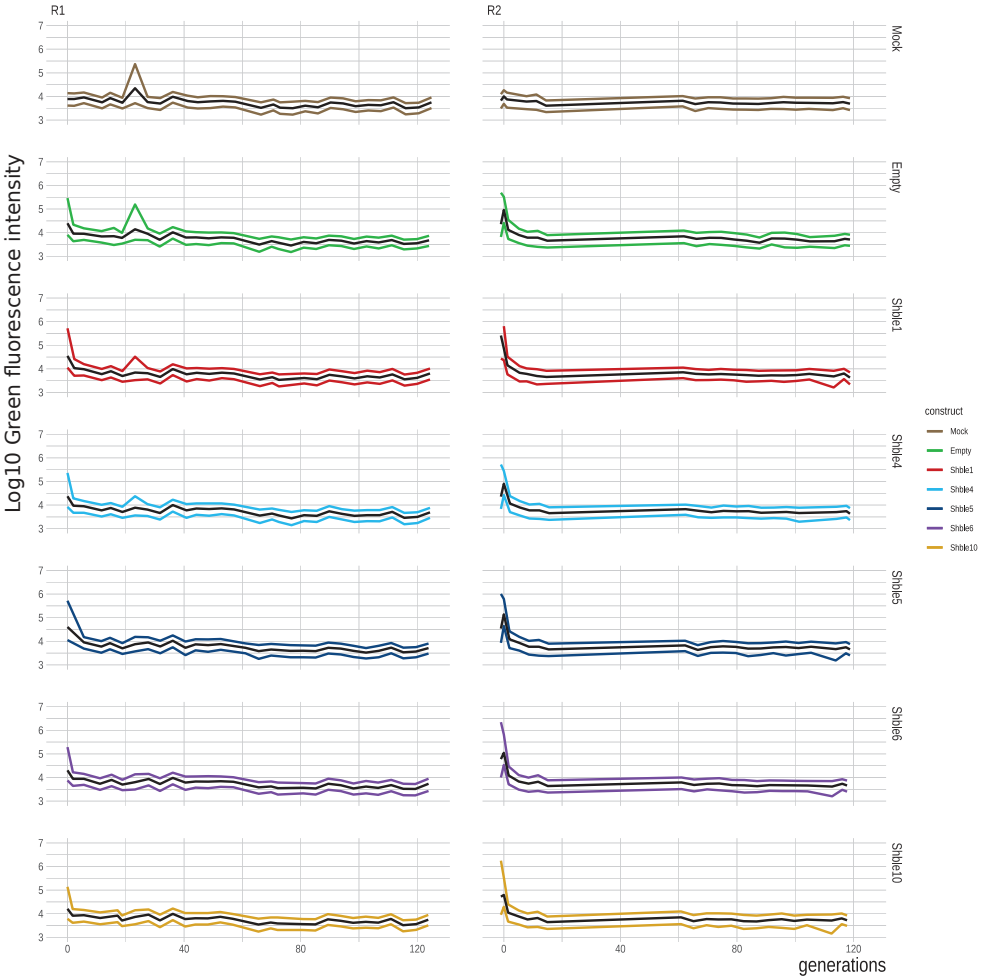


Figure 12: Fluorescence intensity over time – woAB treatment *- Colored lines represent the median of the 20% most (above black line) and least (below black line) fluorescent cells in the population over generations. The Huber central value is represented with a black line. All fluorescence intensity values are in log10. The first column displays R1 and the second R2, while rows display the Constructs.*

Bleomycin treatment

Under the Bleomycin treatment, only the constructs with the *shble* gene survive. The Empty construct only contains the *neo_tp* gene and was eliminated after the first reseeding as the population didn't survive in this treatment. All five constructs showed an overall increase in fluorescence, reaching a plateau after ~10 reseedings in R1 (Figure 13). Despite the CUPrefs differences between the inserted plasmids, there was no notable difference between the average fluorescence intensity of the different constructs for any of the variables followed (Huber value, cumulative fluorescence, 20% most fluorescent cells). As the flow cytometer measures the green fluorescence intensity, size and complexity of each cell individually, we were able to determine if there is within sample heterogeneity in our cell lines (Figure 14). We found the populations homogeneous over time in terms of average fluorescence intensity however when looking at the 20% least fluorescent cells we observe, in Shble1, a seemingly stable sub-population maintaining lower fluorescence values, also visible in the ridge line plots and interval graphs.

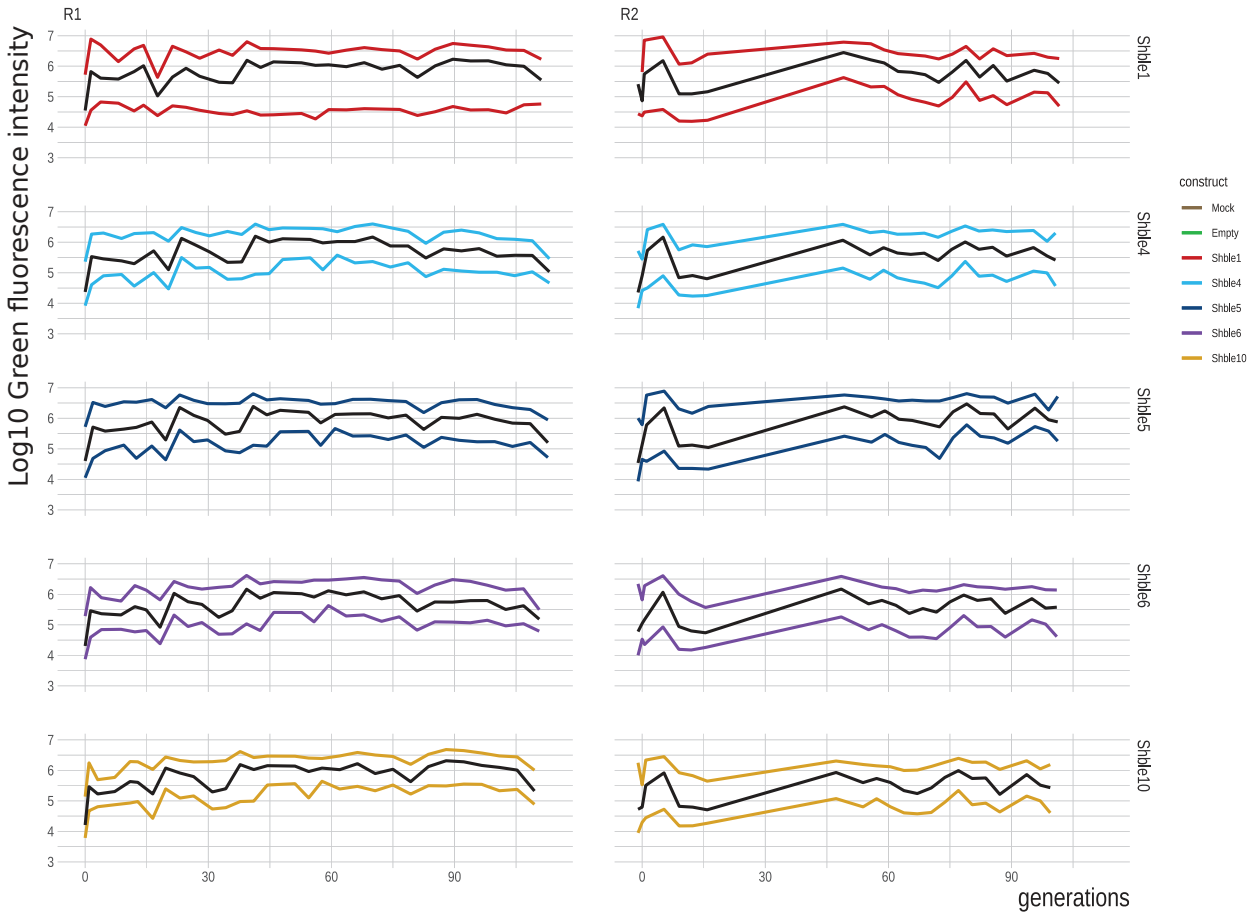
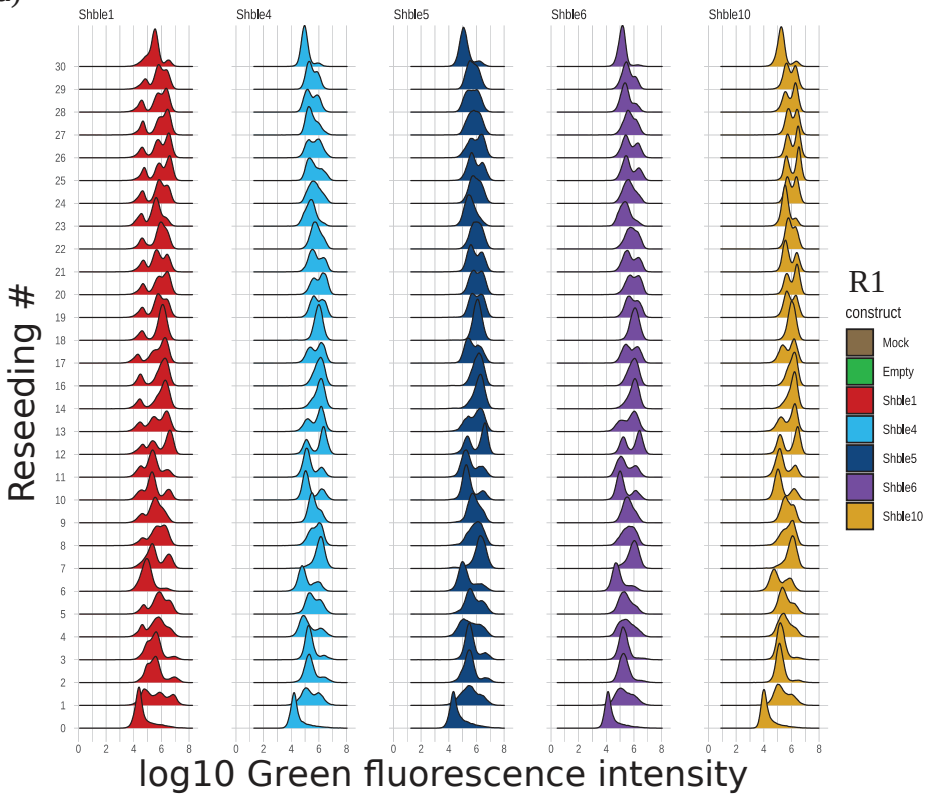


Figure 13: Fluorescence intensity over time – Bleomycin treatment *- Colored lines represent the median of the 20% most (above black line) and least (below black line) fluorescent cells in the population over generations. The Huber central value is represented with a black line. All fluorescence intensity values are in log10. The first column displays R1 and the second R2, while rows display the Constructs.*

a)



b)

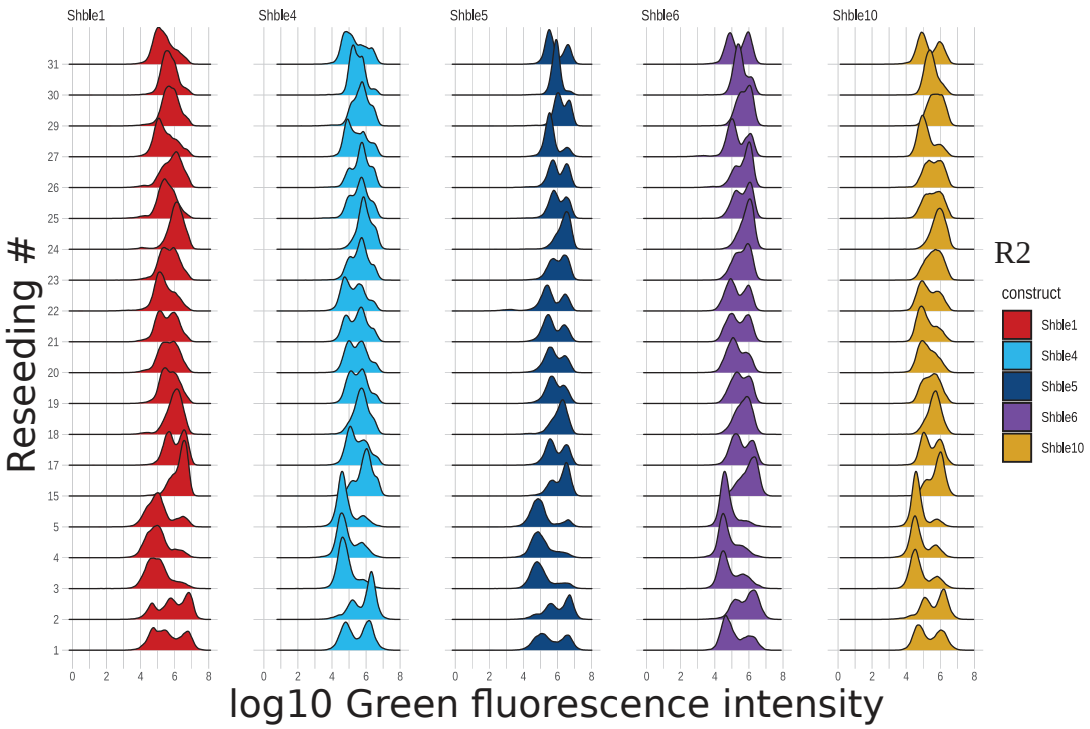


Figure 14: Fluorescence intensity over time – Bleomycin treatment *-Ridge line graph of density of population by fluorescence intensity (x axis – in log10) and by time point (Y axis) for each construct under Bleomycin in a) R1 and b) R2.*

R2 is missing fluorescence values between S5 and S15, but we have the beginning and the end of the second half of the experiment. Under Bleomycin, we see an initial drop of fluorescence at the beginning of the experiment, but at S15 we observe values as high as in R1, although within sample heterogeneity is higher. Cells in the Neomycin treatment, displayed the same initial drop than under Bleomycin, and reached a higher fluorescence by S15. This is not the case however of Shble1, which once again, lost fluorescence. We also observe seemingly stable secondary populations in all the other constructs.

Neomycin treatment

When cells were selected under Neomycin they showed more diversity between constructs than under Bleomycin. We noted a great variability of fluorescence in Shble6 and 10, and a loss of fluorescence in Shble1(Figure 15). In fact Shble1's fluorescence became similar to that of the Mock cell line, albeit a small fraction of the population maintained high values up until 15 Passages. As for Shble6 and 10 these relatively stable secondary populations are were noticeable throughout the whole experiment(Figure 16). In R2, we also detected a loss of fluorescence in Shble1, but much later in the experiment (around S15) compared to the almost immediate loss in R1.

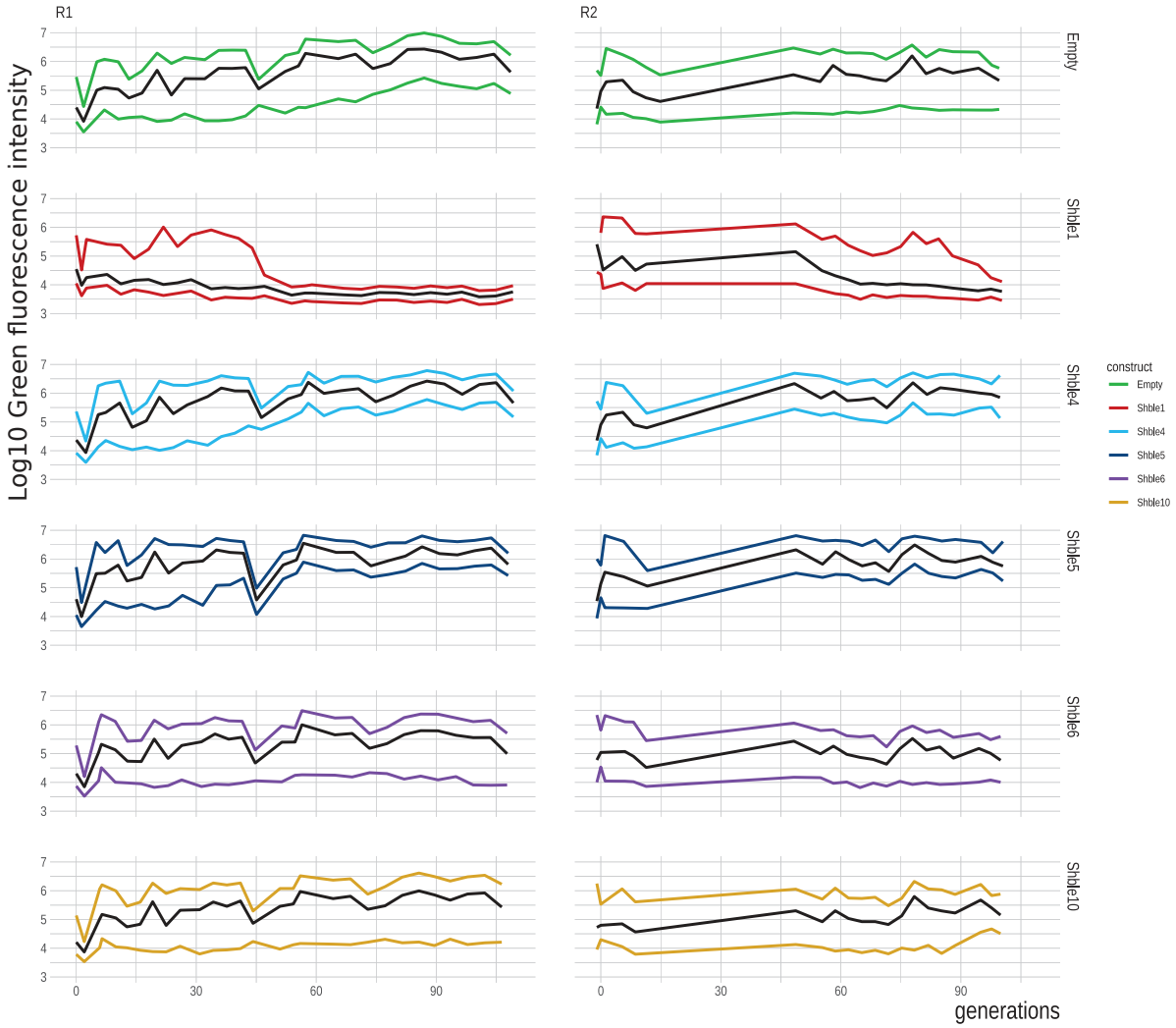


Figure 15: Fluorescence intensity over time – Neomycin treatment *- Colored lines represent the median of the 20% most (above black line) and least (below black line) fluorescent cells in the population over generations. The Huber central value is represented with a black line. All fluorescence intensity values are in log10. The first column displays R1 and the second R2, while rows display the Constructs.*

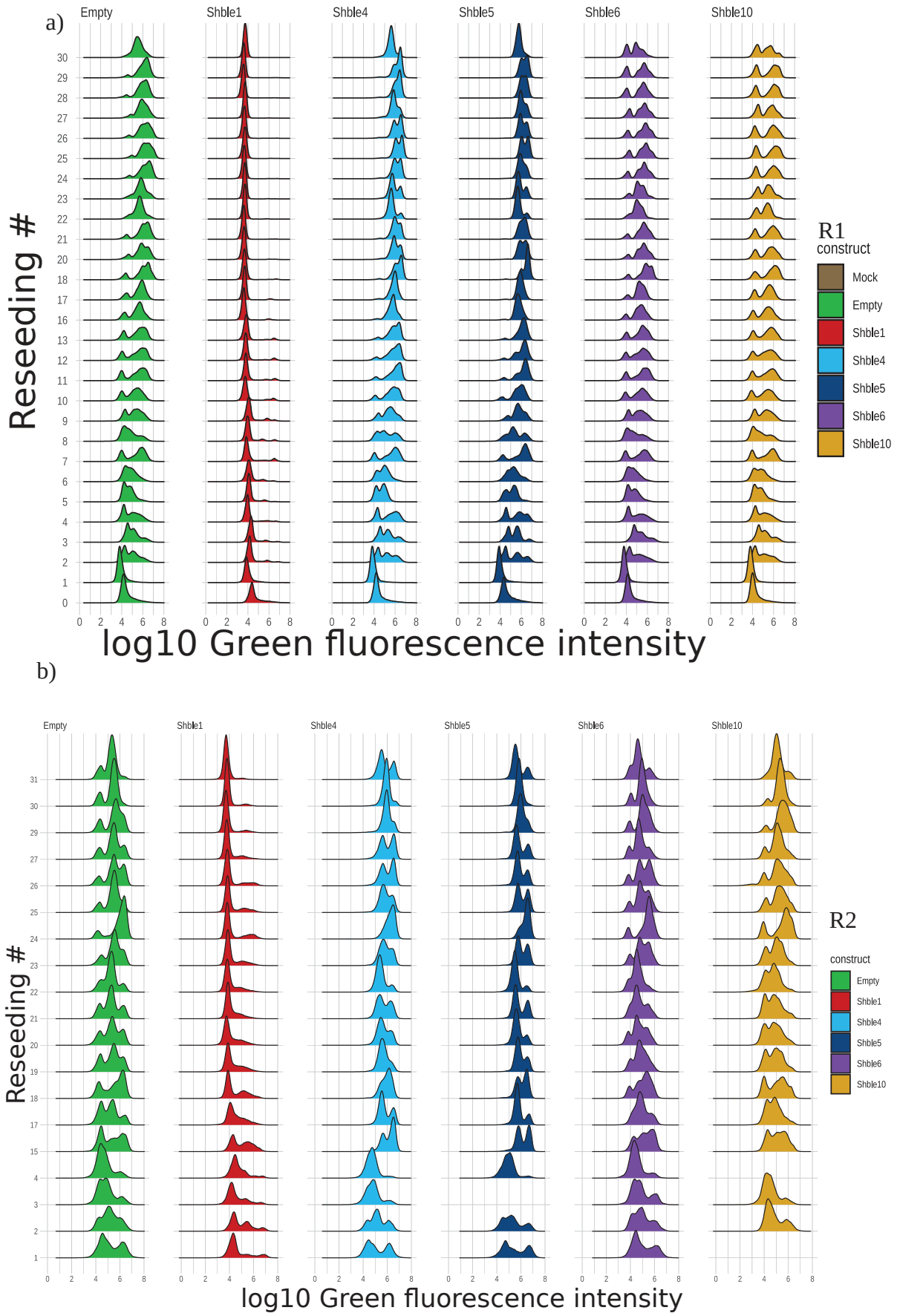


Figure 16: Fluorescence intensity over time – Neomycin treatment *-Ridge line graph of density of population by fluorescence intensity (x axis – in log10) and by time point (Y axis) for each construct under Neomycin in a) R1 and b) R2.*

Comparing the fluorescence levels across constructs and treatments

To further analyze differences between constructs we fitted exponential growth models (or exponential decay, in the case of Shble1 under Neomycin) to the log10 Huber central values, as well as to the median of upper and lower 20% of the cellular population and compared the parameters of the fit. Unfortunately the central value does not capture the multi-modality of the data, but it provides still an idea of the overall picture. With this method we could statistically confirm that the asymptote (highest value over time) of Shble6 is different from that of Shble1,4,5 and Empty. Of course Shble1 is different from all of the other constructs in the same treatment as we cannot fit the same type of model to it due to its loss of fluorescence (exponential decay *vs* exponential growth). We also noted significant differences between Shble10 and the Empty construct (Table 4). Although this approach has the advantage of taking in account the time factor, we were unable to fit models to all of the cell lines. This made us switch to a simpler but more efficient method. We used pairwise comparisons of means between the start of the experiment (S0) and the end of the experiment (last three time-points), assessing the differences with an ANOVA, Pairwise t-test, *post hoc* Tukey test, and effect size analysis (Cohen's d). In virtually all cases, the ANOVA and pairwise t-test showed a significant difference, so in the next paragraphs I will present the overall look of the data, and the effect size, that may be more appropriate to understand the biological significance of the differences than the t-test.

a)

Bleo	lhub	Neo
Shble1	a	Different Model
	i	
	r	
Shble4	a	*
	i	NS
	r	NS
Shble5	a	NS
	i	NS
	r	NS
Shble6	a	NS
	i	NS
	r	NS
Shble10	a	NS
	i	NS
	r	NS

Table 4: Results of comparisons between fitted models to central value over time – we fitted exponential growth models to the curve of evolution of central value of fluorescence over time. For each construct in each cell line, we retrieved the initial value (i), the asymptote(a) and the slope(r) of the fit. These values were then compared by an ANOVA. a) Comparisons of each construct between treatments b) comparisons between each construct in the Bleo treatment c) comparisons between each construct in the Neo treatment. As Shble1 under Neomycin looses fluorescence, an exponential decay model was fitted to it, therefore it is different from the others constructs by default. NS stands for non-significant.

b)

Bleo	lhub	Shble1	Shble4	Shble5	Shble6
Shble4	a	NS	-	-	-
	i	NS	-	-	-
	r	NS	-	-	-
Shble5	a	NS	NS	-	-
	i	NS	NS	-	-
	r	NS	NS	-	-
Shble6	a	*	NS	*	-
	i	NS	NS	NS	-
	r	NS	NS	NS	-
Shble10	a	NS	.	NS	.
	i	NS	NS	NS	NS
	r	NS	NS	NS	NS

c)

Neo	lhub	Shble1	Shble4	Shble5	Shble6	Shble10
Shble4	a	Different Model	-	-	-	-
	i		-	-	-	-
	r		-	-	-	-
Shble5	a		NS	-	-	-
	i		NS	-	-	-
	r		NS	-	-	-
Shble6	a		*	*	-	-
	i		NS	NS	-	-
	r		NS	NS	-	-
Shble10	a		NS	NS	NS	-
	i		NS	NS	NS	-
	r		NS	NS	NS	-
Empty	a		NS	NS	**	*
	i		NS	NS	NS	NS
	r		NS	.	NS	NS

Comparing the mean fluorescence values of the full populations

Overall, results in R1 and in R2 follow the same pattern, albeit with some small notable differences (Table 5). When looking at the starting values, R1 had slightly lower fluorescence intensity values, possibly due to a lower transfection efficiency at the beginning of the experiment. Although the pairwise t-test shows a significant difference between all of the constructs in both replicates, these differences do not exceed medium effect size (Table 6). As for the same differences at the end of the experiment, we will look at them by treatment.

a)

START	Empty		Mock		Shble1		Shble10		Shble4		Shble5	
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
Mock	<2e-16	<2e-16	-	-	-	-	-	-	-	-	-	-
Shble1	<2e-16	0.00012	<2e-16	<2e-16	-	-	-	-	-	-	-	-
Shbl10	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	-	-	-	-	-	-
Shble4	1.5E-10	<2e-16	<2e-16	<2e-16	<2e-16	4.2E-08	<2e-16	<2e-16	-	-	-	-
Shble5	<2e-16	<2e-16	<2e-16	<2e-16	7.6E-13	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	-	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16

b)

END	R1 Bleo			
	Shble1	Shble10	Shble4	Shble5
Shble10	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16

e)

END	R2 Bleo			
	Shble1	Shble10	Shble4	Shble5
Shble10	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16

c)

END	R1 Neo				
	Empty	Shble1	Shble10	Shble4	Shble5
Shble1	<2e-16	-	-	-	-
Shble10	<2e-16	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16

f)

END	R2 Neo				
	Empty	Shble1	Shble10	Shble4	Shble5
Shble1	<2e-16	-	-	-	-
Shble10	<2e-16	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16

d)

END	R1 woAB					
	Empty	Mock	Shble1	Shble10	Shble4	Shble5
Mock	<2e-16	-	-	-	-	-
Shble1	<2e-16	<2e-16	-	-	-	-
Shble10	<2e-16	<2e-16	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16	0.71	<2e-16

g)

END	R2 woAB					
	Empty	Mock	Shble1	Shble10	Shble4	Shble5
Mock	<2e-16	-	-	-	-	-
Shble1	<2e-16	<2e-16	-	-	-	-
Shble10	<2e-16	0.51	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16

Table 5: Pairwise comparisons between the central value of Fluorescence levels at the beginning and at the end of the experiment *- For each construct in each treatment we compared the central value of their fluorescence intensity at the a) beginning of the experiment and at the end of the experiment for each treatment and replicate (b-g). For the beginning of the experiment we took the S0 sample i.e. the time point just before putting cells under selection. For the end of the experiment we concatenated fluorescence results of S26,S27 and S28. Mean comparisons were done by performing an ANOVA and a pairwise t-test, the p-values are displayed in the tables.*

I) Under the Bleo treatment, we observe only small or medium effect sizes, with the exception of Shble4-*vs*-Shble10 in R1.

II) Under Neomycin, we already mentioned that Shble1 loses fluorescence, in effect size, this results in extreme high values. Meanwhile Shble6-*vs*-Shble10 and Shble4-vs-Shble5, we only see a small effect size, hinting to a resemblance between the phenotype of these cell lines.

III) Without antibiotic selection, all cell lines lost fluorescence. Although significant differences are detected between cell lines, it is most likely a difference in auto-fluorescence and not linked to *egfp* expression, therefore it is not relevant to our problematic.

a)

START vs END			
Treatment	Construct	Effect size R1	Effect size R2
Bleo	Shble1	1.455	1.053
	Shble10	3.463	1.007
	Shble4	2.061	1.111
	Shble5	2.047	1.205
	Shble6	2.332	0.619
Neo	Empty	2.290	0.785
	Shble1	2.214	1.242
	Shble10	1.314	0.309
	Shble4	3.007	1.898
	Shble5	3.015	1.379
	Shble6	1.182	0.281
woAB	Empty	2.478	3.017
	Mock	1.186	1.137
	Shble1	2.407	2.981
	Shble10	1.854	2.761
	Shble4	2.395	3.040
	Shble5	2.678	3.868
	Shble6	2.299	3.894

b)

R1	Effect size			
vs	START	END		
		Bleo	Neo	woAB
Sh1-Sh4	0.280	0.362	5.763	0.299
Sh1-Sh5	0.045	0.087	6.848	0.129
Sh1-Sh6	0.390	0.243	2.649	0.303
Sh1-Sh10	0.547	0.382	2.619	0.035
Sh4-Sh5	0.326	0.598	0.150	0.178
Sh4-Sh6	0.109	0.158	1.123	0.002
Sh4-Sh10	0.271	1.037	0.980	0.287
Sh5-Sh6	0.436	0.438	1.293	0.180
Sh5-Sh10	0.593	0.391	1.135	0.103
Sh6-Sh10	0.164	0.861	0.077	0.292

c)

R2	Effect size			
vs	START	END		
		Bleo	Neo	woAB
Sh1-Sh4	0.035	0.107	2.907	0.523
Sh1-Sh5	0.386	0.331	2.862	0.063
Sh1-Sh6	0.230	0.243	1.188	0.495
Sh1-Sh10	0.186	0.368	1.293	0.158
Sh4-Sh5	0.475	0.464	0.126	0.419
Sh4-Sh6	0.294	0.145	1.742	0.144
Sh4-Sh10	0.170	0.270	1.369	0.285
Sh5-Sh6	0.158	0.595	1.678	0.369
Sh5-Sh10	0.620	0.761	1.290	0.097
Sh6-Sh10	0.442	0.110	0.199	0.211

Table 6: Effect Size of Pairwise comparisons between the central value of Fluorescence levels at the beginning and at the end of the experiment *- For each construct in each treatment we compared the central value of their fluorescence intensity at the a) beginning of the experiment and at the end of the experiment for each treatment and replicate (b-c). For the beginning of the experiment we took the S0 sample i.e. the time point just before putting cells under selection. For the end of the experiment we concatenated fluorescence results of S26,S27 and S28. Comparisons were done by performing an effect size analysis (Cohen's d). Effect size is displayed in the table.*

Mean fluorescence values of the 20% most fluorescent cells

To decompose the complexity of our data, we looked at the behavior of the 20% most fluorescent cells. When comparing starting values (S0) R1 and R2 displayed similar patterns, but there's more homogeneity between populations in R2. In fact even the pairwise t-test detected no significant differences in the paired comparisons of Empty-*vs*-Shble10, Shble1-*vs*-Shble5, Shble1-*vs*-Shble6 and Shble5-*vs*-Shble6. (Table 7).

a)

START	Empty		Mock		Shble1		Shble10		Shble4		Shble5	
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
Mock	<2e-16	<2E-016	-	-	-	-	-	-	-	-	-	-
Shble1	<2e-16	<2E-016	<2e-16	<2E-016	-	-	-	-	-	-	-	-
Shbl10	<2e-16	0.96	<2e-16	<2E-016	<2e-16	<2E-016	-	-	-	-	-	-
Shble4	<2e-16	6.00E-09	<2e-16	<2E-016	<2e-16	<2E-016	<2e-16	2.20E-06	-	-	-	-
Shble5	<2e-16	<2E-016	<2e-16	<2E-016	0.0011	0.96	<2e-16	<2E-016	<2e-16	<2E-016	-	-
Shble6	<2e-16	<2E-016	<2e-16	<2E-016	<2e-16	0.96	<2e-16	<2E-016	<2e-16	<2E-016	<2e-16	0.99

b)

END	R1 Bleo			
	Shble1	Shble10	Shble4	Shble5
Shble10	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16

e)

END	R2 Bleo			
	Shble1	Shble10	Shble4	Shble5
Shble10	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16

c)

END	R1 Neo				
	Empty	Shble1	Shble10	Shble4	Shble5
Shble1	<2e-16	-	-	-	-
Shble10	<2e-16	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16

f)

END	R2 Neo				
	Empty	Shble1	Shble10	Shble4	Shble5
Shble1	<2e-16	-	-	-	-
Shble10	<2e-16	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	<2e-16	-	-
Shble5	<2e-16	<2e-16	<2e-16	0.055	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16

d)

END	R1 woAB					
	Empty	Mock	Shble1	Shble10	Shble4	Shble5
Mock	<2e-16	-	-	-	-	-
Shble1	<2e-16	<2e-16	-	-	-	-
Shble10	<2e-16	<2e-16	<2e-16	-	-	-
Shble4	1.90E-05	<2e-16	<2e-16	<2e-16	-	-
Shble5	<2e-16	3.80E-06	<2e-16	3.00E-15	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16	1.90E-05	<2e-16

g)

END	R2 woAB					
	Empty	Mock	Shble1	Shble10	Shble4	Shble5
Mock	<2e-16	-	-	-	-	-
Shble1	<2e-16	<2e-16	-	-	-	-
Shble10	1	<2e-16	<2e-16	-	-	-
Shble4	<2e-16	<2e-16	<2e-16	<2e-16	-	-
Shble5	1	<2e-16	<2e-16	1	<2e-16	-
Shble6	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16

Table 7: Pairwise comparisons between the upper values of Fluorescence levels at the beginning and at the end of the experiment *- For each construct in each treatment we compared the mean fluorescence value of the 20% most fluorescent cells at the a) beginning of the experiment and at the end of the experiment for each treatment and replicate (b-g). For the beginning of the experiment we took the S0 sample i.e. the time point just before putting cells under selection. For the end of the experiment we concatenated fluorescence results of S26,S27 and S28. Mean comparisons were done by performing an ANOVA and a pairwise t-test, the p-values are displayed in the tables.*

At the end of the experiment, under Bleomycin, the one notable difference between R1 and R2 was the higher fluorescence intensity in R1- Shble10. What stays consistent between the two replicates is the low effect size of the paired differences of Shble1-*vs*-Shble5, Shble4-*vs*-Shble6 and the very large effect size of the paired differences in Shble1-*vs*-Shble6 (Table 8).

Under Neomycin what immediately jumps to the eye, is the much higher variance of Shble1 in R2 despite keeping the same pattern otherwise. We observe a small effect size for the paired comparison Shble4-*vs*-Shble5, and a large effect size between Shble6 and Shble10.

a)

START vs END			
Treatment	Construct	Effect size	
		R1	R2
Bleo	Shble1	2.523	1.348
	Shble10	4.077	1.330
	Shble4	1.855	2.109
	Shble5	1.961	1.687
	Shble6	2.521	0.791
Neo	Empty	3.063	2.108
	Shble1	3.357	1.082
	Shble10	3.266	0.824
	Shble4	3.079	3.204
	Shble5	2.637	2.102
	Shble6	2.352	0.539
woAB	Empty	5.753	3.856
	Mock	2.229	3.256
	Shble1	4.176	5.890
	Shble10	3.808	4.262
	Shble4	4.641	4.575
	Shble5	5.121	5.014
	Shble6	4.972	6.570

b)

R1	Effect size			
vs	START	END		
		Bleo	Neo	woAB
Sh1-Sh4	0.476	2.268	6.795	0.418
Sh1-Sh5	0.049	0.769	7.065	0.214
Sh1-Sh6	0.669	1.733	5.633	0.414
Sh1-Sh10	0.880	0.394	6.088	0.174
Sh4-Sh5	0.533	1.394	0.317	0.211
Sh4-Sh6	0.170	0.437	1.733	0.054
Sh4-Sh10	0.375	2.029	0.866	0.306
Sh5-Sh6	0.734	0.922	2.092	0.194
Sh5-Sh10	0.950	0.462	1.209	0.061
Sh6-Sh10	0.214	1.478	0.853	0.303

c)

R2	Effect size			
vs	START	END		
		Bleo	Neo	woAB
Sh1-Sh4	0.593	0.459	3.076	0.807
Sh1-Sh5	0.017	0.374	2.952	0.135
Sh1-Sh6	0.016	1.310	0.887	1.107
Sh1-Sh10	0.517	1.286	1.413	0.137
Sh4-Sh5	0.597	0.882	0.050	0.437
Sh4-Sh6	0.596	0.946	4.773	0.316
Sh4-Sh10	0.073	0.958	3.286	0.439
Sh5-Sh6	0.000	1.802	4.351	0.635
Sh5-Sh10	0.518	1.685	3.010	0.002
Sh6-Sh10	0.518	0.231	0.959	0.634

Table 8: Effect Size of Pairwise comparisons between the upper values of Fluorescence levels at the beginning and at the end of the experiment *- For each construct in each treatment we compared the mean fluorescence value of the 20% most fluorescent cells at the a) beginning of the experiment and at the end of the experiment for each treatment and replicate (b-c). For the beginning of the experiment we took the S0 sample i.e. the time point just before putting cells under selection. For the end of the experiment we concatenated fluorescence results of S26,S27 and S28. Comparisons were done by performing an effect size analysis (Cohen's d). Effect size is displayed in the table.*

Comparing the mean fluorescence values of the 20% least fluorescent cells

At the start of the experiment, R2 displayed higher values of fluorescence compared to the least fluorescent cells of R1 which were in fact not fluorescent at all, or very close to the threshold of auto-fluorescence. Under the Bleomycin treatment we observe a clear difference between R1 and R2, while R1-Shble1 displayed a very low almost non-fluorescent value, R2 presented much higher overall values, especially for Shble5.

The least fluorescent cells under Neo treatment displayed the same behavior in R1 and R2: Shble1 did not present fluorescent cells, Shble6 and Shble10 were just at the brink of fluorescence (4.2), and even the least fluorescent cells of Shble4 and Shble5 showed around 5.5 fluorescence intensity. Effect size analysis shows that Shble6-*vs*-Shble10 and Shble4-*vs*-Shble5 have actually negligible differences in both replicates, while the rest of the pairwise analysis shows very large effect size (Table 9).

a)

START vs END			
Treatment	Construct	Effect size	
		R1	R2
Bleo	Shble1	1.564	1.347
	Shble10	6.041	1.675
	Shble4	5.733	1.370
	Shble5	5.000	1.685
	Shble6	5.354	0.286
Neo	Empty	2.850	0.088
	Shble1	3.895	4.795
	Shble10	1.687	0.891
	Shble4	3.816	2.509
	Shble5	6.723	1.894
	Shble6	0.756	1.716
woAB	Empty	3.831	3.029
	Mock	1.438	1.058
	Shble1	3.958	4.261
	Shble10	2.431	2.622
	Shble4	3.540	3.087
	Shble5	4.221	4.748
	Shble6	3.554	4.791

b)

R1	Effect size			
vs	START	END		
		Bleo	Neo	woAB
Sh1-Sh4	0.699	1.777	6.029	0.374
Sh1-Sh5	0.035	2.346	10.514	0.146
Sh1-Sh6	0.999	1.894	3.071	0.516
Sh1-Sh10	1.552	2.970	3.759	0.108
Sh4-Sh5	0.788	0.896	0.639	0.233
Sh4-Sh6	0.323	0.293	3.752	0.128
Sh4-Sh10	0.953	1.797	3.430	0.463
Sh5-Sh6	1.116	0.554	6.221	0.372
Sh5-Sh10	1.722	0.952	5.891	0.248
Sh6-Sh10	0.648	1.445	0.450	0.601

c)

R2	Effect size			
vs	START	END		
		Bleo	Neo	woAB
Sh1-Sh4	0.145	0.048	6.151	1.060
Sh1-Sh5	1.142	1.017	6.012	0.209
Sh1-Sh6	0.579	0.353	1.923	0.527
Sh1-Sh10	0.409	0.174	0.979	0.483
Sh4-Sh5	0.844	1.201	0.240	0.906
Sh4-Sh6	0.372	0.340	3.708	0.755
Sh4-Sh10	0.463	0.145	2.998	0.456
Sh5-Sh6	0.465	1.152	3.726	0.291
Sh5-Sh10	1.407	1.323	3.050	0.338
Sh6-Sh10	0.871	0.243	0.221	0.155

Table 9: Effect Size of Pairwise comparisons between the lower values of Fluorescence levels at the beginning and at the end of the experiment *- For each construct in each treatment we compared the mean fluorescence value of the 20% least fluorescent cells at the a) beginning of the experiment and at the end of the experiment for each treatment and replicate (b-c). For the beginning of the experiment we took the S0 sample i.e. the time point just before putting cells under selection. For the end of the experiment we concatenated fluorescence results of S26,S27 and S28. Comparisons were done by performing an effect size analysis (Cohen's d). Effect size is displayed in the table.*

Results for DNA and mRNA levels

To assess plasmid DNA levels changes in our cells, we performed qPCRs on two targets (Figure 10), both of them located on the *shble-egfp* complex present in the inserted transfected plasmids. The correlation between the results obtained for the two targets is good (0.67 and 0.78 Spearman's rank correlation) in R1 and R2 respectively. Thus it is possible to pool the values of the two targets to follow plasmid level changes, making our final value the mean of four measures (two targets with a duplicate each). The two replicates (R1 & R2) however cannot be pooled as despite their similarities, we see differences in their behavior and it is best to analyze them separately (Figure 17).

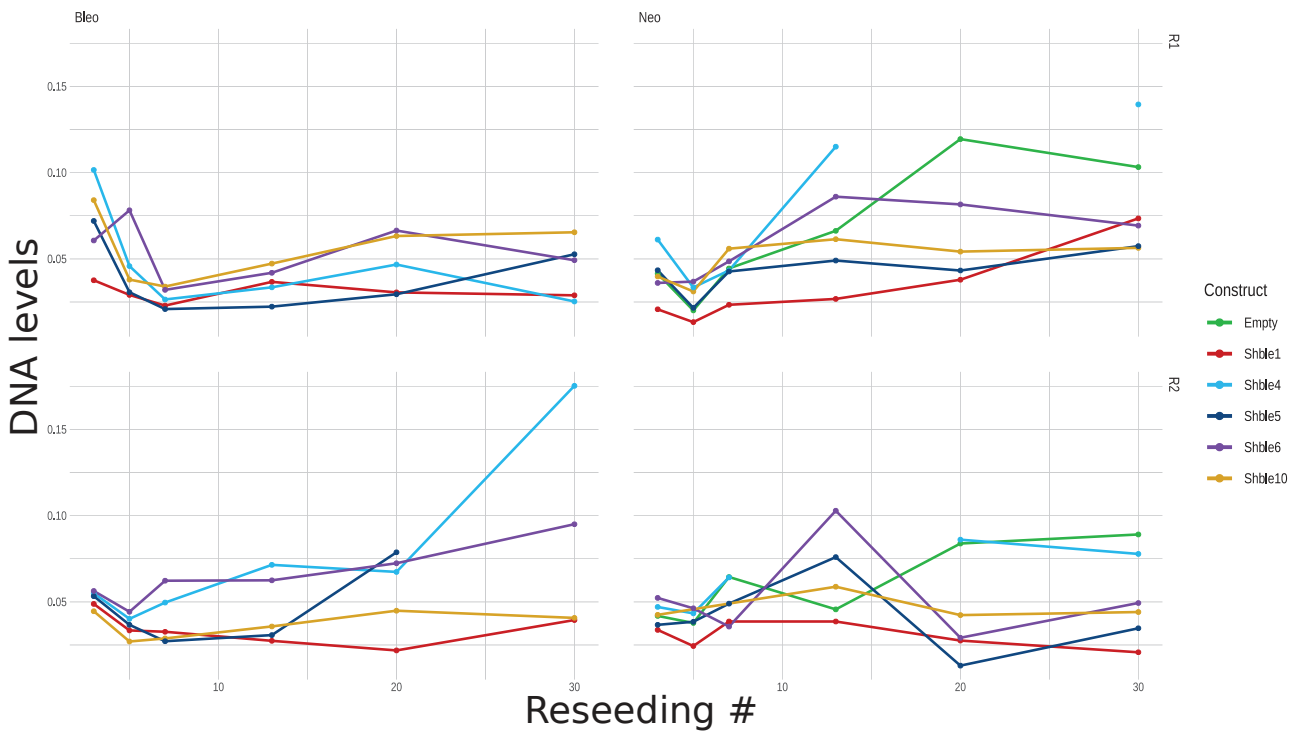


Figure 17: DNA levels over time *– qPCR results of R1 and R2 for each construct and each treatment. In y-axis the reseeding number i.e. sampling time-points, in x-axis the mean 2-ΔΔCT of the 2 primer pairs (5UTR, P2A)*

To study mRNA level changes in our cells the rt-qPCR data were prepared in the same way, as for the qPCR (Figure 18). This was possible because the targets (AU1 -P2A) are found on the same *shble-egfp* mRNA molecule, and the correlation between targets is 0.93 for both R1 and R2.

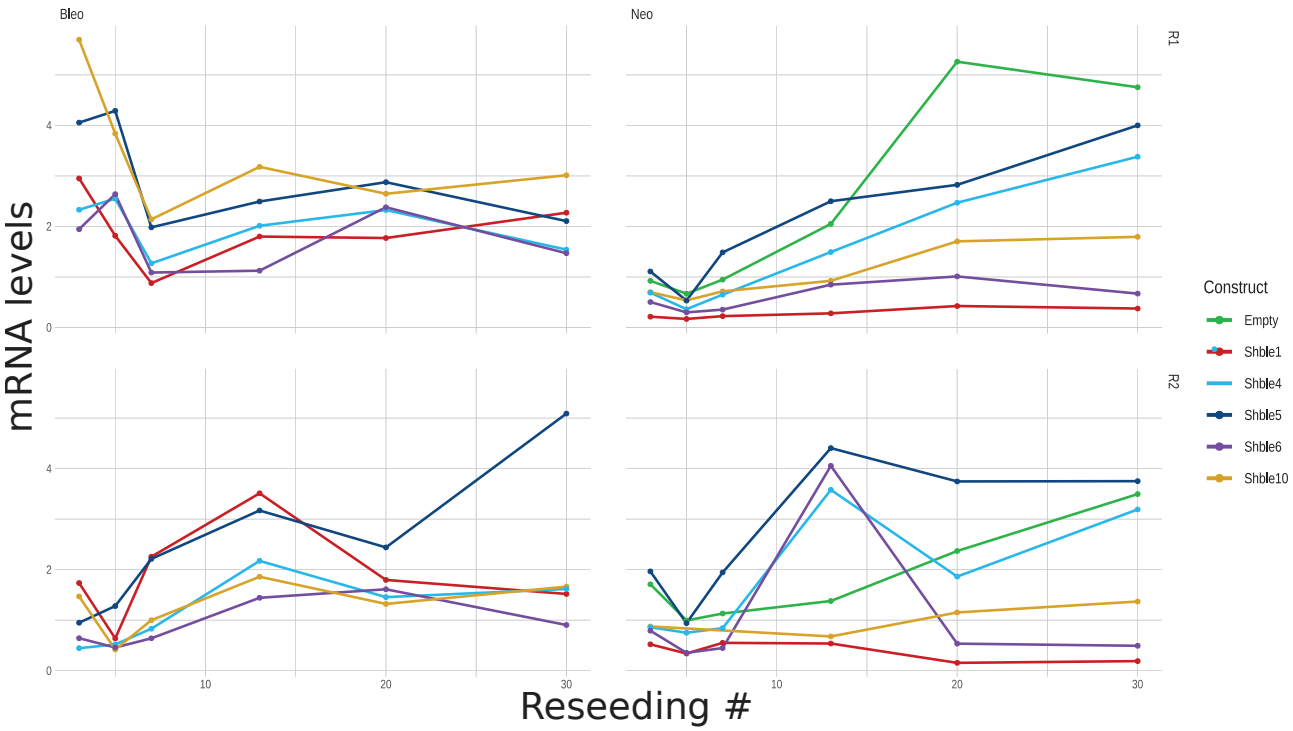


Figure 18: mRNA levels over time *– rt-qPCR results of R1 and R2 for each construct and each treatment. In y-axis the reseeding number i.e. sampling time-points, in x-axis the mean 2-ΔΔCT of the 2 primer pairs (5UTR, P2A)*

Analyses for DNA levels through time and selection

With a one-way ANOVA, we checked if plasmid levels changed significantly over time. In R1 under Bleomycin we do not detect such change (Table 10), but it is worth to note the initial drop of values between S3 and S5. Under Neomycin however Shble1, Shble4 and the Empty construct did change over time. All three construct showed increasing values, but Shble1 was much lower than the other two mentioned construct.

Plasmid Levels	pr(>F)			
	Neo - Constructs		Bleo- Constructs	
	R1	R2	R1	R2
Sh1	**0.00417**	0.248	0.785	0.616
Sh4	**0.0512**	**0.0781**	0.279	**0.0192**
Sh5	0.141	0.608	0.983	**0.0985**
Sh6	0.125	0.929	0.791	**0.00649**
Sh10	0.221	0.867	0.691	0.398
Empty	**0.0223**	**0.0266**	-	-

Table 10: ANOVA of DNA levels over time *– p-values of a one-way ANOVA test, to see if the slope of DNA levels is different from 0.*

Despite the same protocol, R2 shows some dissimilarities from R1, most importantly the starting values were much more closer to each other between constructs than in the first replicate. Under Bleomycin, Shble4 Shble5 and Shble6 display a significant change over time, and have increasing plasmid levels, that were overall higher than that of Shble1 and Shble10. Cells in the Neomycin treatment showed a more similar pattern to R1, as both Shble4 and the empty construct display significant change over time. Indeed these two constructs has increasing values over time, and a much higher final value than the other constructs.

Analyses for mRNA levels through time and selection using rt-qPCR

Once again we used a one-way ANOVA, this time, to detect the effect of time on mRNA levels in our cell populations (Table 11). In R1 under Bleomycin, just as the plasmid levels mRNA levels do not change significantly over time, apart from the drop between S3 and S5(which is not captured by the ANOVA). Meanwhile, cell lines in the Neomycin treatment all have significant p-values, except Shble6. Shble1, has only a slight increase, Shble10 a mild one, while Shble4, Shble5 and Empty are rapidly increasing over the other constructs.

In R2, only Shble5 under Bleomycin and Shble1 and Empty under Neomycin have a significant p-value. Even if the ANOVA doesn't detect it as significant, we can see the same tendencies under Neomycin in R2 than in R1, where Shble4, Shble5 and Empty have much higher mRNA values than other construct, and Shble10 positioned in the middle, lower than the aforementioned constructs, but higher than Shble1 and Shble6.

mRNA Levels	pr(>F)			
	Neo - Constructs		Bleo- Constructs	
	R1	R2	R1	R2
Sh1	**0.023**	**0.092**	0.926	0.927
Sh4	**0.000**	0.102	0.509	0.139
Sh5	**0.002**	0.103	0.206	**0.011**
Sh6	0.196	0.786	0.848	0.291
Sh10	**0.004**	0.198	0.342	0.309
Empty	**0.011**	**0.014**	-	-

Table 11: ANOVA of mRNA levels over time *– p-values of a one-way ANOVA test, to see if the slope of mRNA levels is different from 0.*

Analyses for mRNA levels through time and selection using RNASeq

To quantify and analyze the transcriptome, we recovered RNAseq data via the genewiz platform. In total 44 samples were sequenced for R1 (time-point 3, 7,13, 28), 33 samples for R2 (3,7,30), and a time-point 0, before selection for each Construct. We identified 19812 host genes, from which 18480 have been identified in at least four samples, and we could also quantify the presence of the heterologous transcript *shble-egfp* and of *neo_tp* the genes present on the inserted plasmid.

We identified alternative spliced forms of Shble4 and Shbl6, although these had not been predicted by any algorithms. Shble4 has one alternative spliced form, while Shble6 has two. In both replicates and treatments the full forms stays dominant, except in R2 Shble6 under Neomycin at S30, where we detected a drastic amount -over ten times more- spliced forms than the full form (Figure 20). In Shble4, the spliced forms only represent 4% (mean of all measures) of the construct mRNA, while in Shble6 they are more present : 30%, (mean excluding the outliers – S0, S30,.)

While analyzing the mRNA sequences for mutations, we unexpectedly found a loss of coverage in R1 Shble1 under Neomycin (Figure 19). In fact the end of the mRNA (from around 1015 bp-1115 bp – just around the P2A), which contains the *egfp* part of the *shble-egfp* part is being less and less detected in the population over time. We found no mutations that would explain this loss, moreover, all the mutations found (Table 12) are present in such low frequencies, that they should not have a visible effect when analyzing the full populations. We identified 31 mutations using breseq, from which ten were identified as false-positive. Curiously mutations found in the *shble* sequence appear under Neomycin selection, and vice-versa, mutations in the *neo_tp* sequence appeared under the Bleomycin treatment. The frequency of these mutations, besides being very low, also doesn't seem vary over time.

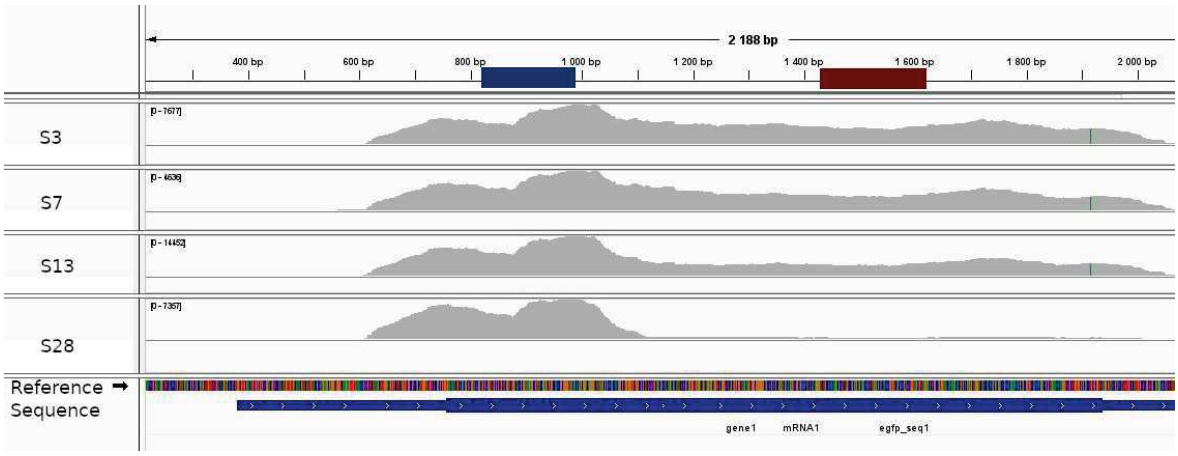


Figure 19: Loss of coverage in R1 Shble1 under Neo treatment *– Output of the IGV software of Shble1 R1 under Neomycin over time. rt-qPCR primers would attach where they are marked in blue (5UTR primer) and red (P2A primer). Coverage of the egfp gene is lost over time.*

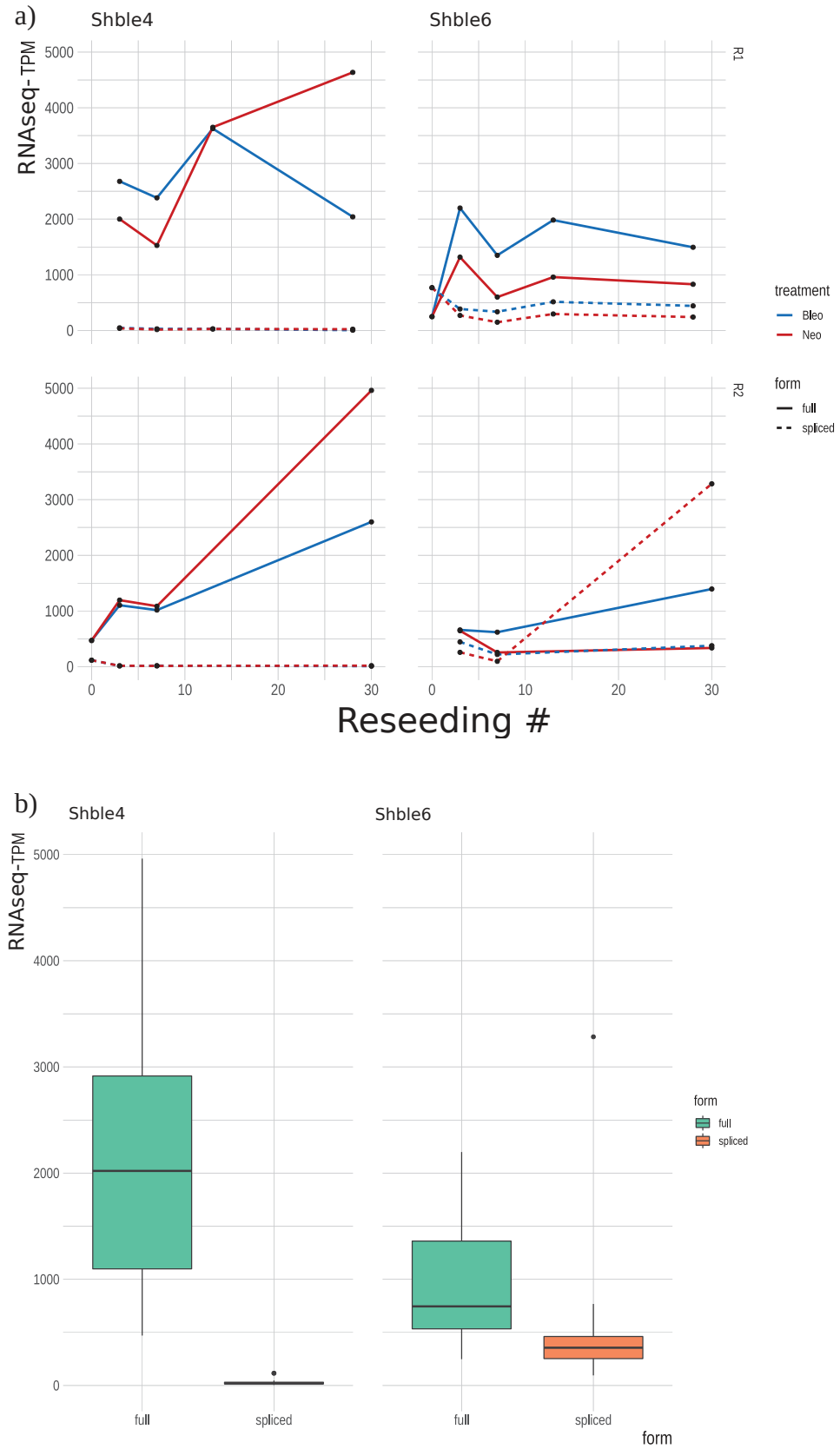


Figure 20: Splicing in Shble4 and Shble6 *– a) RNAseq results (TPM) of full and spliced forms of Shble4 and Shble6 a) by treatment and by replicate over time. b) boxplot of overall TPM condensing all time-points, treatments and replicates.*

Construct	Time-point	Selection	Gene	Position #	Genome position	Test(F&K-S)	Mutation	Frequency	Change type	Amino acid	Type
Shble1	S13	Neo	Shble-egfp	1266	Shble-CDS	NS	C>T	7.60%	F92F(TTC→TTT)	Phe→Phe	Synonymous
Shble1	S28	Neo	Shble-egfp	1432	P2A-CDS	NS	1 bp deletion	9.50%	coding(51/77nt)		
Shble1	S28	Neo	Shble-egfp	1434	P2A-CDS	NS	A>T	9.60%	E18V(GAG→GTG)	Glu→Val	Non-Synonymous
Shble1	S28	Neo	Shble-egfp	1467	egfp-CDS	NS	G>T	9.40%	K3N(AAG→AAT)	Lys→Asp	Non-Synonymous
Shble1	S28	Neo	Shble-egfp	1476	egfp-CDS	NS	G>A	10.50%	E6E(GAG→GAA)	Glu→Glu	Synonymous
Shble1	S28	Neo	Shble-egfp	1486	egfp-CDS	NS	G>A	10.60%	G10R(GGG→AGG)	Gly→Arg	Non-Synonymous
Shble4	S28	Neo	Shble-egfp	2116	egfp-CDS	NS	C>T	12.40%	L220L(GGG→TTG)	Leu→Leu	Synonymous
Shble4	S28	Neo	Shble-egfp	2198	3'UTR_shble	NS	C>T	11.10%	intergenic		
Shble4	S28	Neo	Shble-egfp	2204	3'UTR_shble	NS	C>T	10.90%	intergenic		
Shble4	S28	Neo	Shble-egfp	2216	3'UTR_shble	NS	C>T	10.90%	intergenic		
Shble4	S28	Neo	Shble-egfp	2231	3'UTR_shble	NS	C>T	11.20%	intergenic		
Shble4	S28	Neo	Shble-egfp	2254	3'UTR_shble	*(F)	C>A	14.20%	intergenic		
Shble4	S28	Neo	Shble-egfp	2258	3'UTR_shble	*(F)	C>A	13.90%	intergenic		
Shble6	S7	Bleo	Neo-tp	3320	neomycine-CDDS_neo	*(K-S)	G>A	5.30%	M1I(ATG→ATA)	Met→Iso	Non-Synonymous
Shble6	S7	Bleo	Neo-tp	3324	neomycine-CDDS_neo	*(K-S)	G>A	5.20%	E3K(GAA→AAA)	Glu→Lys	Non-Synonymous
Shble6	S13	Bleo	Neo-tp	3320	neomycine-CDDS_neo	NS	G>A	7.40%	M1I(ATG→ATA)	Met→Iso	Non-Synonymous
Shble6	S13	Bleo	Neo-tp	3324	neomycine-CDDS_neo	NS	G>A	7.30%	E3K(GAA→AAA)	Glu→Lys	Non-Synonymous
Shble6	S13	Bleo	Neo-tp	3330	neomycine-CDDS_neo	NS	G>A	7.90%	D5N(GAT-AAT)	Asp→Asn	Non-Synonymous
Shble6	S13	Bleo	Neo-tp	3398	neomycine-CDDS_neo	NS	G>C	8.20%	Q27H(CAG→CAC)	Gln→His	Non-Synonymous
Shble6	S28	Bleo	Neo-tp	3320	neomycine-CDDS_neo	NS	G>A	7.00%	M1I(ATG→ATA)	Met→Iso	Non-Synonymous
Shble6	S28	Bleo	Neo-tp	3324	neomycine-CDDS_neo	NS	G>A	6.90%	E3K(GAA→AAA)	Glu→Lys	Non-Synonymous
Shble6	S28	Bleo	Neo-tp	3330	neomycine-CDDS_neo	NS	G>A	6.80%	D5N(GAT-AAT)	Asp→Asn	Non-Synonymous
Shble6	S28	Bleo	Neo-tp	3398	neomycine-CDDS_neo	NS	G>C	6.20%	Q27H(CAG→CAC)	Gls→His	Non-Synonymous
Shble6	S13	Neo	Neo-tp	3301	5'UTR_neo	NS	G>C	5.10%	intergenic		
Shble6	S28	Neo	Neo-tp	3301	5'UTR_neo	NS	G>C	9.60%	intergenic		
Shble1	S3	Bleo	Shble-egfp	1915	egfp-CDS	*(F)	A>C	5.80%	M153L(ATG→CTG)	Met→Leu	Non-Synonymous
Shble4	S3	Bleo	Shble-egfp	1915	egfp-CDS	*(F)	A>C	5.90%	M153L(ATG→CTG)	Met→Leu	Non-Synonymous
Shble6	S3	Bleo	Shble-egfp	1915	egfp-CDS	*(F)	A>C	6.30%	M153L(ATG→CTG)	Met→Leu	Non-Synonymous
Shble1	S7	Bleo	Shble-egfp	1915	egfp-CDS	*(F)	A>C	5.90%	M153L(ATG→CTG)	Met→Leu	Non-Synonymous
Shble5	S7	Bleo	Shble-egfp	1915	egfp-CDS	*(F)	A>C	5.30%	M153L(ATG→CTG)	Met→Leu	Non-Synonymous
Shble..	S..	Bleo/Neo	Shble-egfp	2150	egfp-CDS	*(F)	T>A	100.00%	L231H(CTC→CAC	Leu→His	Non-Synonymous

Table 12: Results of the mutational analyses obtained by breseq *- All the mutations obtained are presented here. The first three columns represent the evolutionary conditions in which a change of base was observed. The lines highlighted in red represent mutations for which at least one of the two tests (Fisher (F) or Kolmogorov-Smirnov (K-S)) stands out as significant (*). The line highlighted in green represents the same thing same thing but this time for a fixed mutation (frequency = 100%).*

Overall in R1 we have detected more RNA than in R2 (Figure 21). In R1 only Shble4 and Empty under Neomycin treatment shows a detectable time effect after running an ANOVA. Shble6 have the lowest TPM values while Shble4 has the highest values in both Bleo and Neo treatment, Shble1 under Neomycin also has very low TPM values. When we analyze the data of R1 via an ANOVA followed by a *post hoc* Tukey test, we see a significant difference between the two treatments (p adj. : 0.002). Indeed, both Shble1 and Shble10 have significant p-values, if we compare the two treatment. Now, if we ignore the time and treatment factor and take in account only the differences between construct, we detect a significant difference between Shble6 versus Shble4, Shble5, Shble10,Empty, and between Shble1 versus Shble5 (Figure 22). Under the Bleomycin treatment only Shble6 versus Shble10 has a detectable and significant difference. Under Neomycin however, Shble1 is significantly different from Shble4, Shble5 and Empty, while Shble6 is significantly different from Shble5 and Empty.

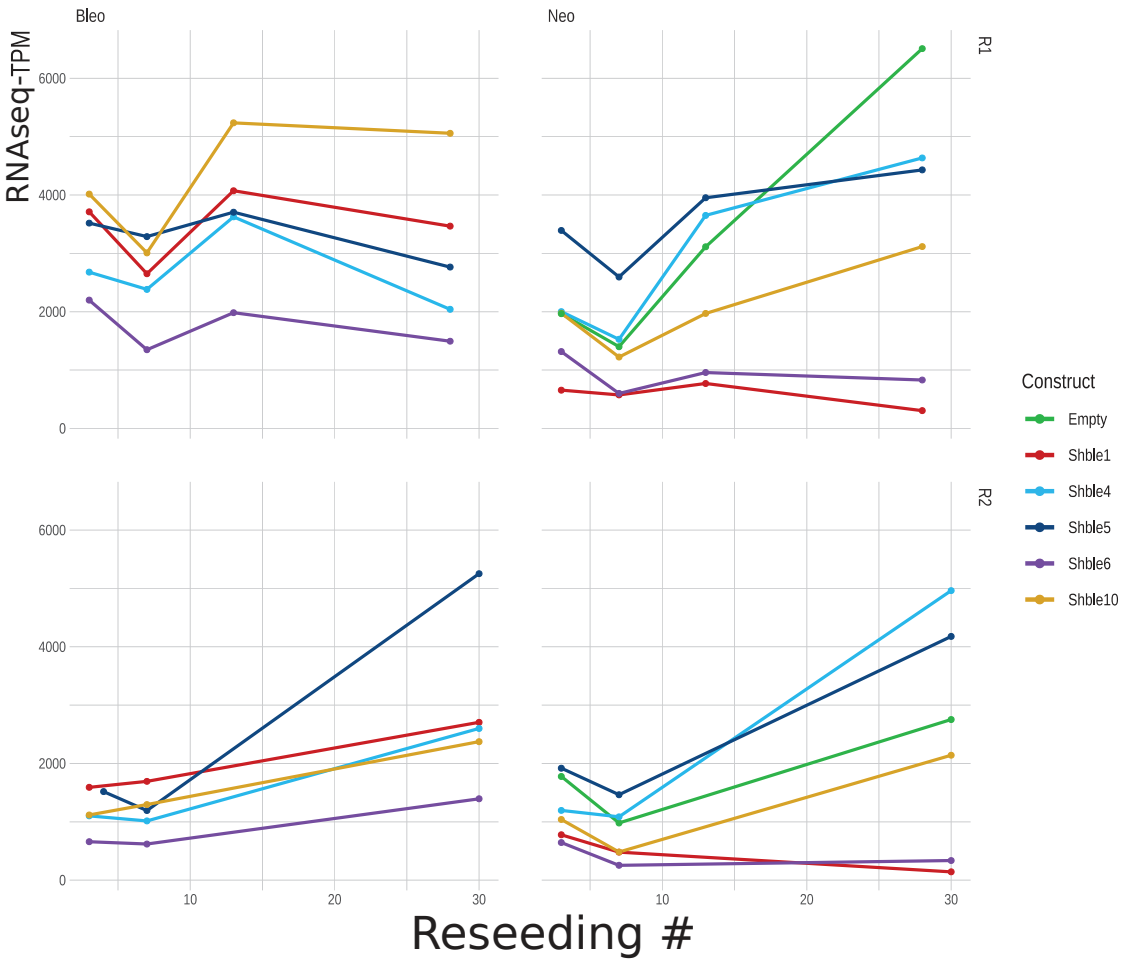


Figure 21: RNA levels over time - RNAseq *– RNAseq results of R1 and R2 for each construct and each treatment over time. In y-axis the reseeding number i.e. sampling time-points, in x-axis the TPM.*

In R2, we observe a similar pattern with R1 under Neomycin. Shble4 and Shble5 has very high TPM values, Shble10 has a medium value, and Shble1 and Shble6 has TPM values close to zero. Under Bleomycin we observe increasing TPM values for all construct, Shble6 having lower values than the other constructs.

As in R2 we have only three time points, the statistical tests should be taken with a grain of salt, nevertheless we do detect a significant change over time in Shble1 and Shble10 under Bleomycin.

When not taking in account the treatment nor the time factor, nothing is significant apart from a difference between Shble5 and Shble6 (p adj = 0.091), Shble5 having one of the highest TPM values, while Shble6 has the lowest ones.

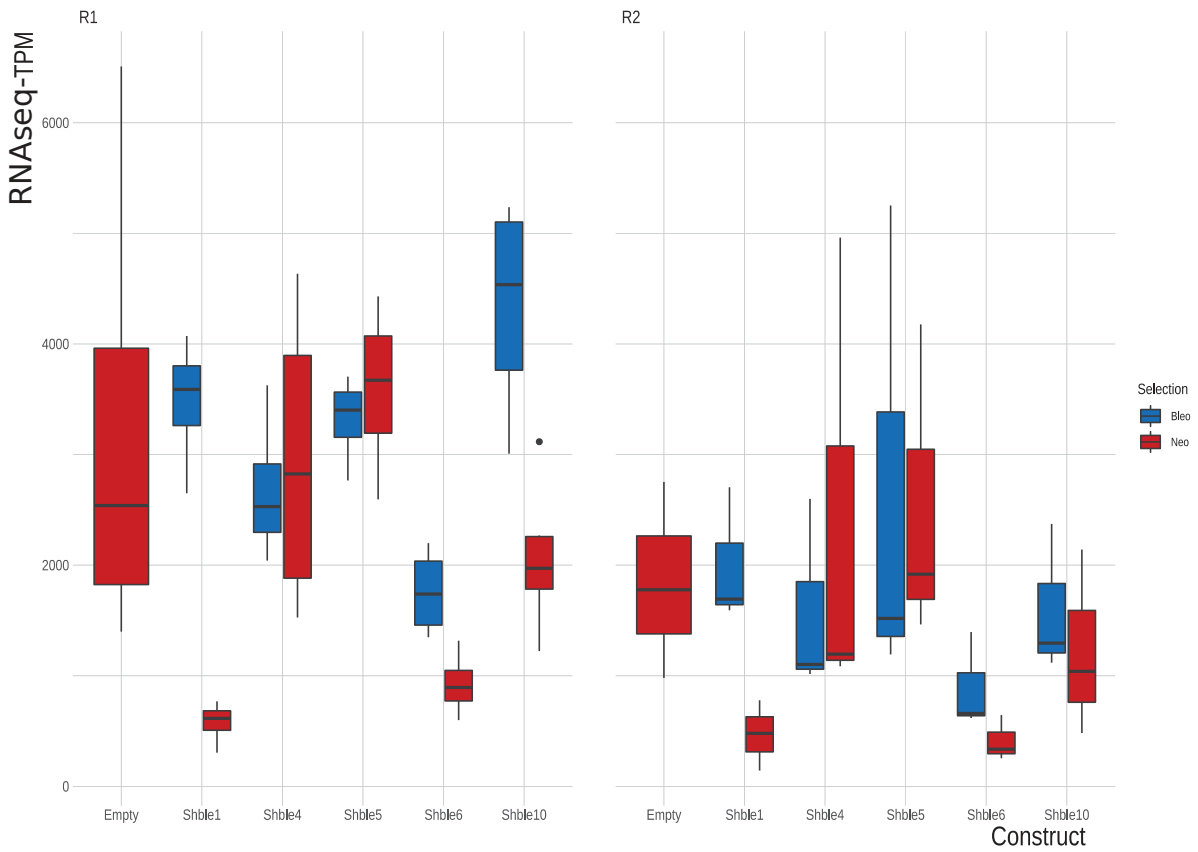


Figure 22: Overall RNA levels - RNAseq *– Boxplot of RNAseq results of R1 and R2 for each construct and each treatment. .*

Analyses for Protein levels through time and selection using mass spectrometry

3309 different proteins were identified in the samples of R1, two of these proteins are the heterologous proteins expressed from the inserted plasmids, the rest: cellular proteins. Curiously, NEO_TP, the protein that confers resistance against Neomycin is not detected in any of our samples despite its mRNA being detected in high values by RNAseq. The potential SHBLE* proteins translated from the spliced forms of Shble4 and Shble6 were also missing from the proteome. We observe a high correlation between the protein levels detected for SHBLE and EGFP (overall value 0.9 rho Spearman's rank correlation, p-value=1.281e-14) (Figure 23). It is also notable that although in theory SHBLE and EGFP should be translated from the same mRNA and thus be observed in the same quantity, we detect three times higher levels for EGFP than for SHBLE.

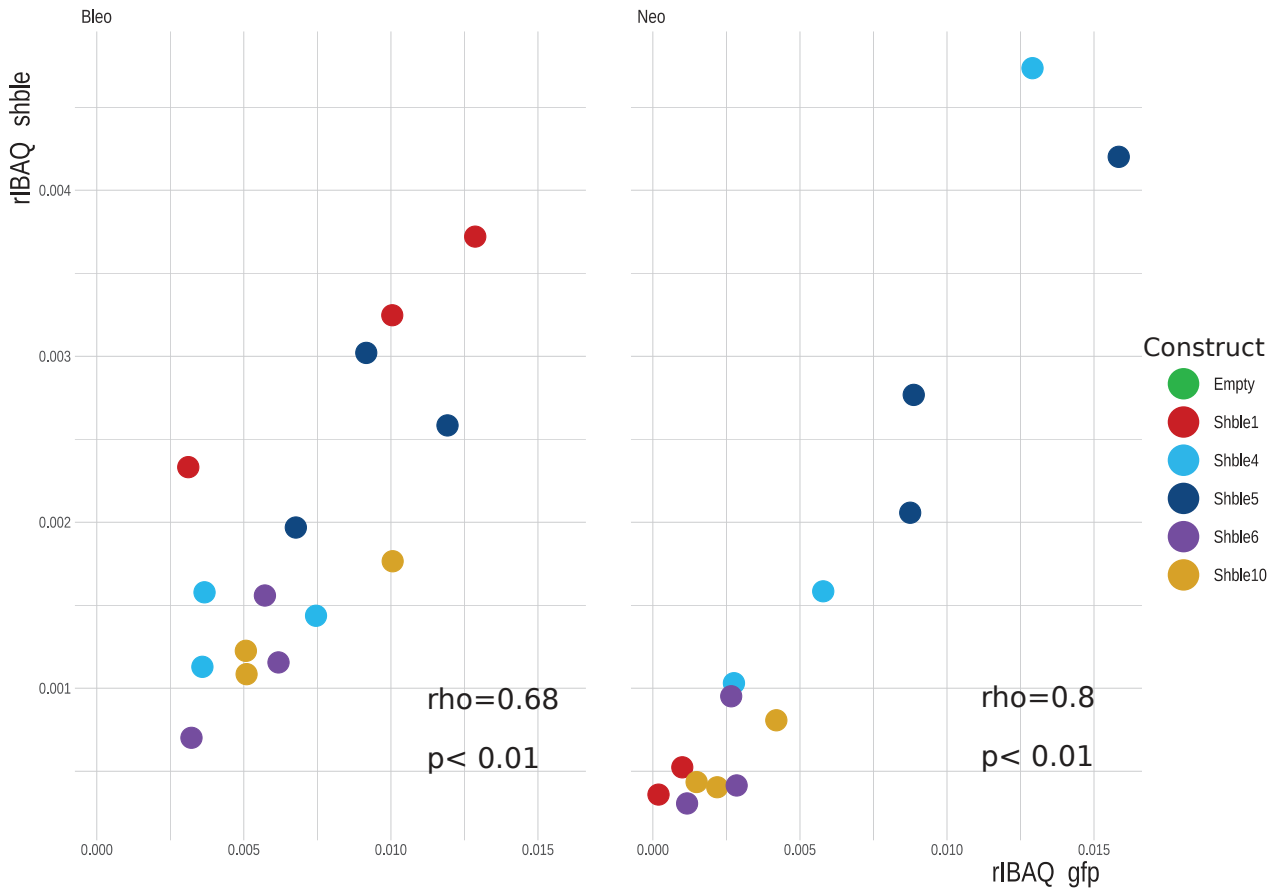


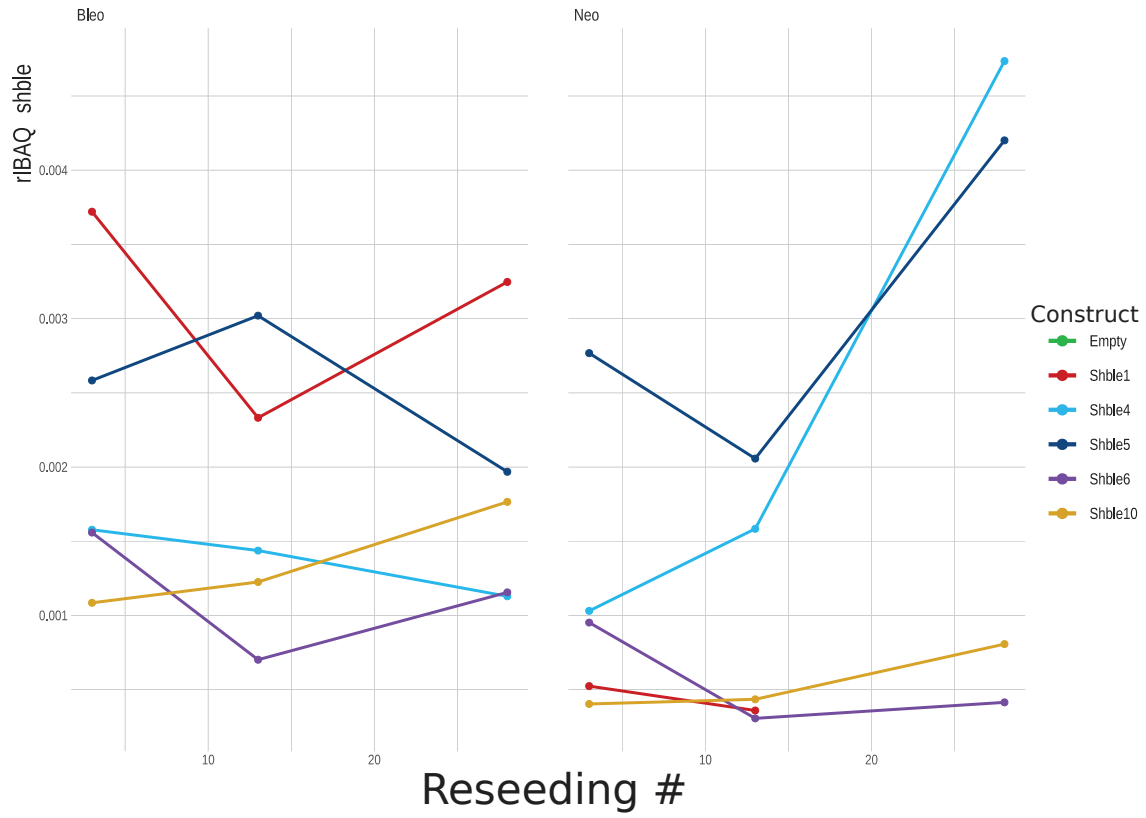
Figure 23: SHBLE-EGFP protein level correlation *– Correlation between the rIBAQ of EGFP and SHBLE in all R1 samples by treatment. Correlation was calculated by Spearman's rank correlation.*

We performed a one-way ANOVA to see if time has an effect but as we have only three time-points by cell line, the p-values might not be accurate (Table 13. Therefore we will talk about observed trends, rather than statistically solid results when describing cell lines individually. When looking at evolution of SHBLE levels over time under Bleomycin, we observed that Shble1 and Shble5 have overall higher rIBAQ values than Shble4, Shble6 and Shble10. This values are rather stable over time. Under the Neomycin treatment however Shble4 and Shble5 have values that increase drastically over time while Shble1, Shble6 and Shble10 have low values throughout the experiment, with Shble1 not being detected at the last time-point probably due to too low concentration (Figure 24).

Protein Levels	pr(>F)			
	Neo		Bleo	
	GFP	SHBLE	GFP	SHBLE
Sh1	0.315	-	0.892	0.855
Sh4	0.072	0.172	0.916	0.0615
Sh5	0.269	0.471	0.0994	0.531
Sh6	0.863	0.507	0.834	0.834
Sh10	0.42	0.216	0.263	0.135
Empty	0.251	-	-	-

Table 13: ANOVA of protein levels over time – p-values of a one-way ANOVA test, to see if the slope of protein levels is different from 0.

a)



b)

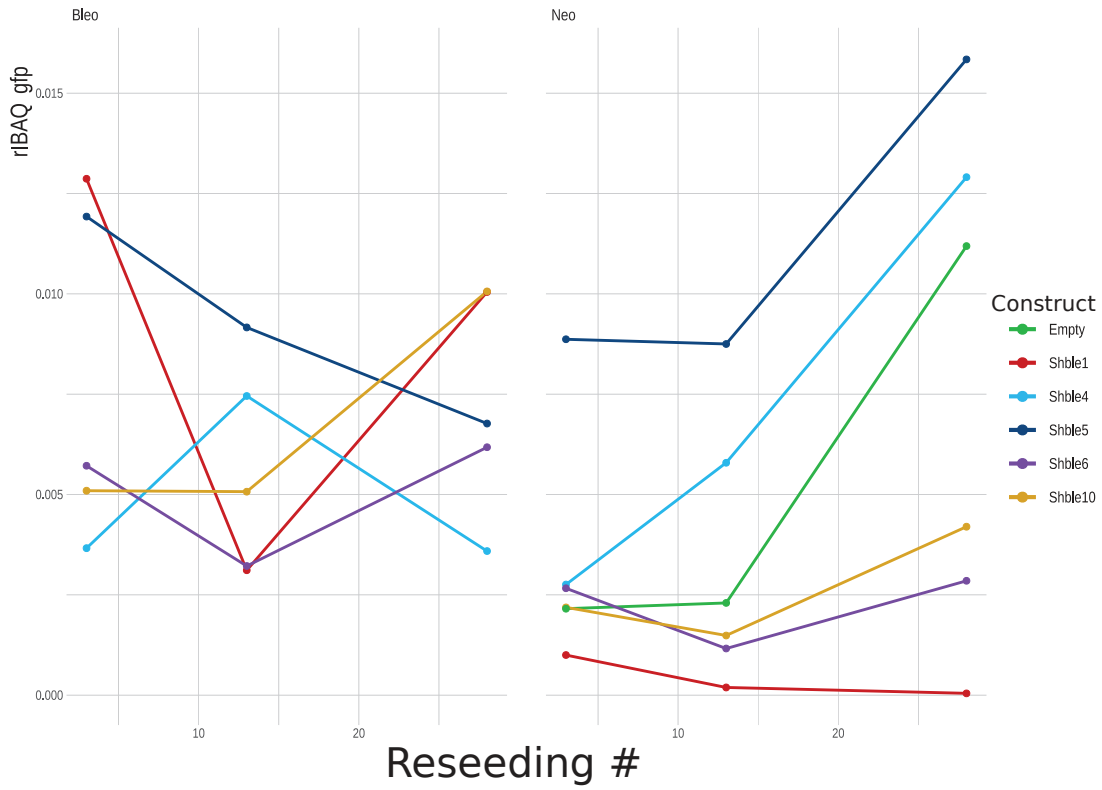


Figure 24: Protein levels over time *– Mass spectrometry results of R1 for each construct and each treatment over time. In y-axis the reseeding number i.e. sampling time-points, in x-axis the rIBAQ of a) SHBLE and b) EGFP. .*

If we ignore the time factor, we can compare the effect of the treatments or the different constructs on translation, to do so we performed a two-way ANOVA followed by a *post hoc* Tukey test. Overall, if we take in consideration all protein levels of a construct without distinguishing between treatments and time-points we detect a significant difference between Shble5 and Shble10 and between Shble5 and Shble6 in both SHBLE and EGFP levels, and between Shble5 and Shble1 in EGFP levels (Table 14). The treatment only has a statistically detectable effect on EGFP (p adj 0.094), when not considering time and construct, but we do see a significant effect between Shble1 Bleo-Neo in SHBLE levels. Now if we look at cell lines individually there are no differences in neither SHBLE nor EGFP levels between any cell lines under Bleomycin, but under Neomycin, Shble5 is significantly different from Shble1, Shbl6 and Shble10 concerning SHBLE levels, and Shble1 and 6 concerning EGFP levels (Figure 25).

SHBLE	diff	lwr	upr	p adj
Shble1 Neo-Bleo	-0.0027	-0.0053	0.0000	0.0545
Shble10 Neo-Bleo	-0.0008	-0.0032	0.0016	0.9637
Shble4 Neo-Bleo	0.0011	-0.0013	0.0035	0.8412
Shble5 Neo-Bleo	0.0005	-0.0019	0.0029	0.9990
Shble6 Neo-Bleo	-0.0006	-0.0030	0.0018	0.9961
EGFP	**diff**	**lwr**	**upr**	**p adj**
Shble1 Neo-Bleo	-0.0083	-0.0181	0.0016	0.1587
Shble10 Neo-Bleo	0.0022	-0.0076	0.0121	0.9992
Shble4 Neo-Bleo	0.0019	-0.0080	0.0117	0.9999
Shble5 Neo-Bleo	-0.0028	-0.0127	0.0070	0.9946
Shble6 Neo-Bleo	-0.0041	-0.0140	0.0057	0.9191

Table 14: Comparison of overall protein levels *– Results of a post hoc Tukey test between protein levels of each construct in the Neo treatment against itself in the Bleo treatment.*

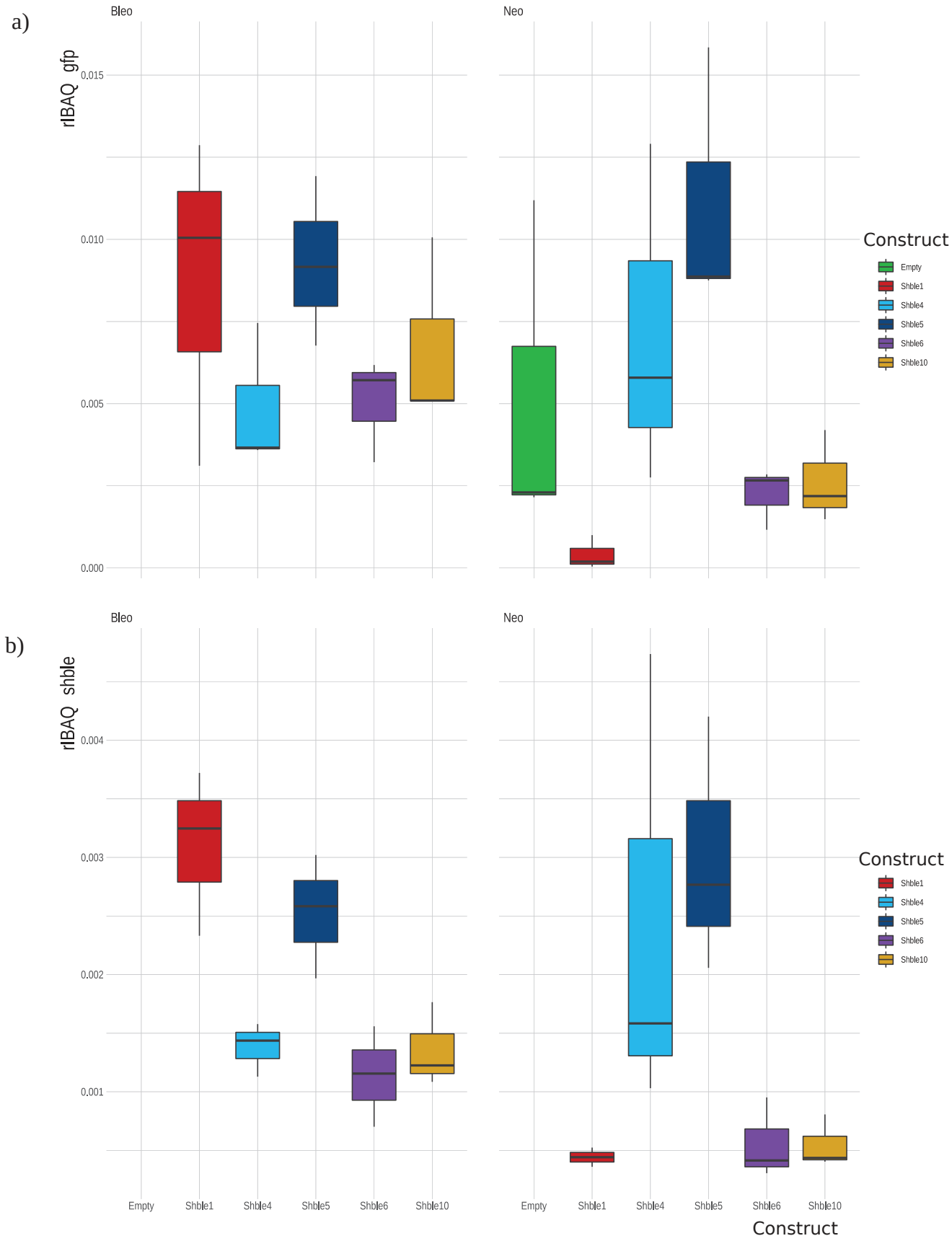


Figure 25: Overall Protein levels *– Boxplot of mass spectrometry results of R1 for each construct and each treatment. a) rIBAQ of EGFP and b) rIBAQ of SHBLE*

Correlations between molecular steps from genotype to phenotypes

DNA levels (qPCR) *vs* mRNA levels (rt-qPCR)

Once we inspected all steps of gene expression individually, we looked at the links and correlations between the different molecular species during the information flow process. Overall, variation in the DNA levels were not good predictors of variation in mRNA levels (Figure 26). We used Spearman's rank correlation test to assess correlation between plasmid DNA levels, and mRNA levels detected by qPCR and rt-qPCR respectively. In R1 we observed an overall 0.39 correlation (rho, p-value=0.001), but if we look at the correlation by treatment, the Neomycin treatment has much higher correlation (rho = 0.70, p-value 4.047e-06) compared to the Bleomycin treatment (rho = 0.45, p-value = 0.01145).

In R2 the overall correlation is lower (rho=0.257, p-value = 0.042), and under Neomycin we detected a correlation of 0.489 (p-value = 0.004). Under Bleomycin there was no significant correlation observed in R2.

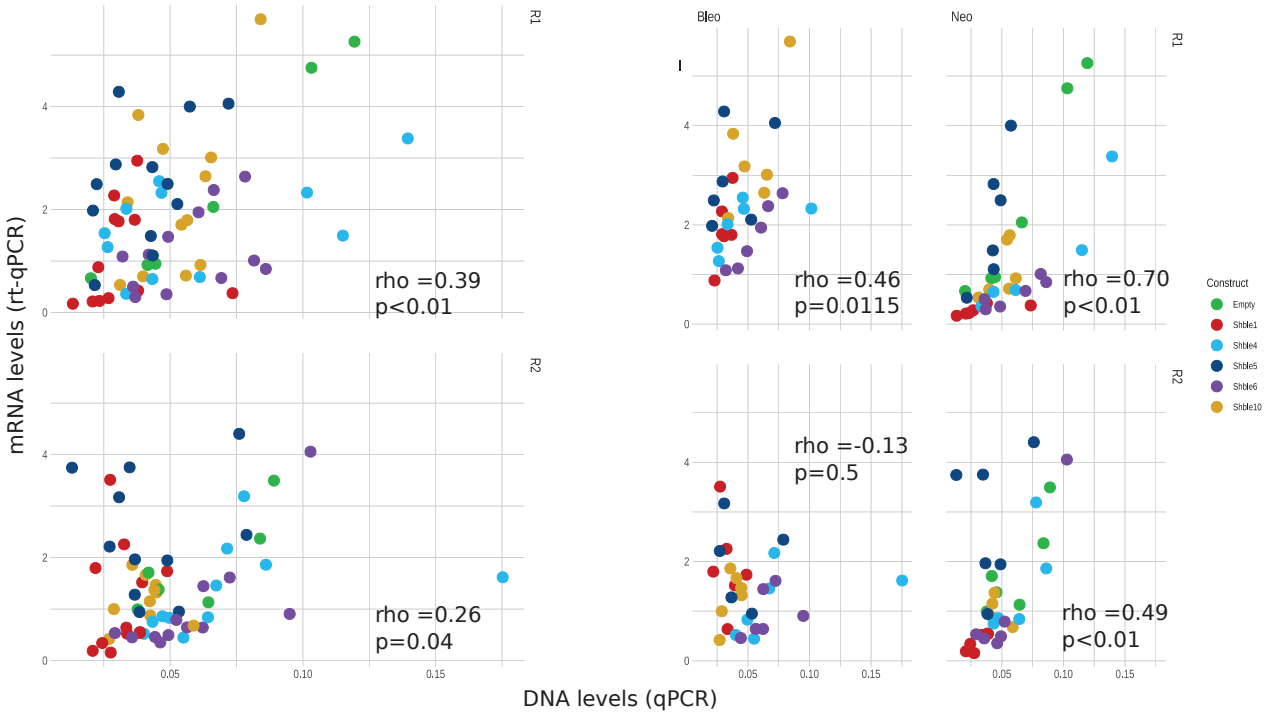


Figure 26: Correlation between DNA and mRNA levels *– We calculated the correlation between DNA and mRNA levels by using Spearman's rank correlation. Results are colored by Construct, and grouped by treatment and by replicate. The first panel is the overall correlation, and the second panel is the correlation when considering treatments separately.*

Targeted mRNA levels (rt-qPCR) *vs* overall mRNA levels (RNAseq)

We assessed correlation between the detected mRNA levels for the same targets, when detected by rt-qPCR and by RNAseq, as this allows us to confirm the results of RNAseq (Figure 27). Overall we observe a very high Spearman correlation of 0.879 (p-value = 4.38e-07) and 0.87 (p-value = 3.05e-07) in R1 and R2 respectively, between the two method. But when stratifying by treatments, Neomycin displayed a very good correlation of 0.925 (p-value < 2.2e-16) in R1 and 0.946 (p-value < 2.2e-16) in R2, while Bleomycin is lower at 0.707 (p-value = 0.004) for R1 and 0.758 (p-value = 0.003) for R2.

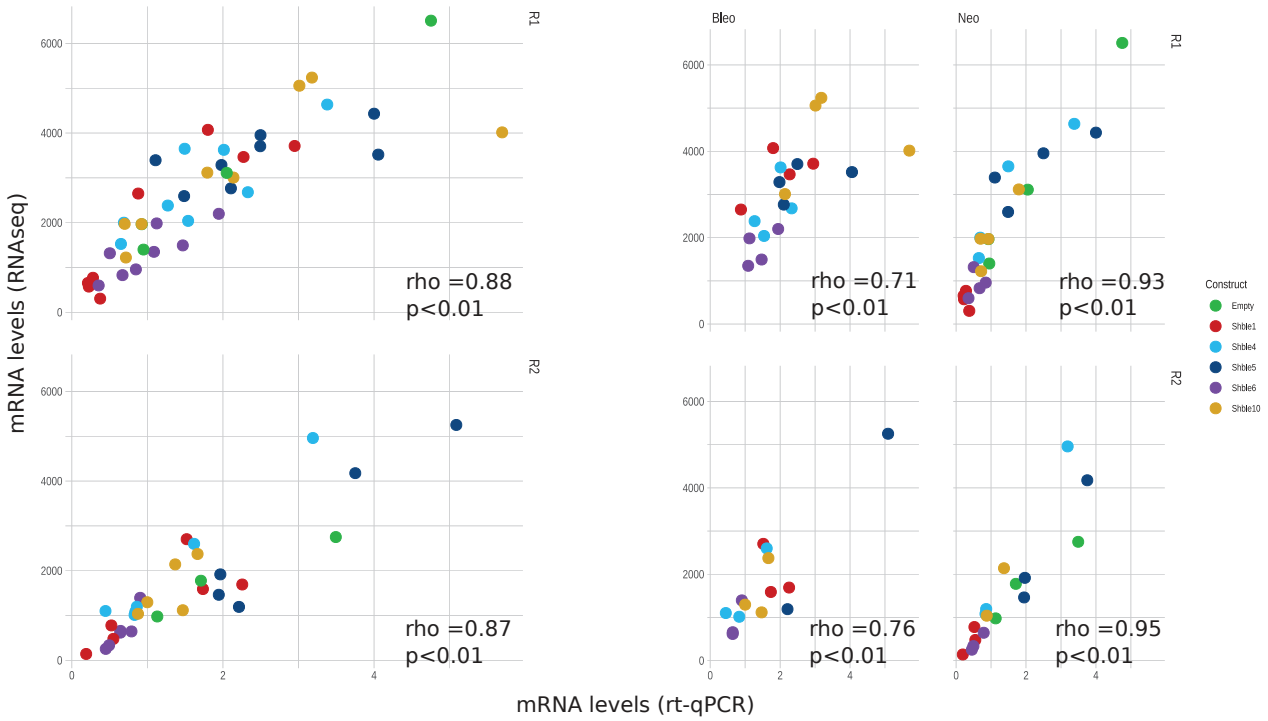


Figure 27: Correlation between mRNA levels measured by rt-qPCR vs RNAseq*– We calculated the correlation between rt-qPCR and RNAseq results by using Spearman's rank correlation. In each graph results are colored by Construct, and grouped by treatment and by replicate. The first panel is always the overall correlation, and the second panel is the correlation when considering treatments separately.*

mRNA levels *vs* Protein levels

Although we didn't detect all the proteins for which we observed the corresponding mRNA, we could still verify the correlation between mRNA and protein levels for SHBLE and EGFP in R1 (Figure 28). Variations in the overall transcriptome levels of the *shble-egfp* mRNA complex showed a 0.80 of correlation versus the protein levels of EGFP (p-value= 3.728e-10), and a 0.76 correlation

versus the protein levels of SHBLE (p-value = 2.759e-08). When stratifying by treatments, the correlation with mRNA levels is once again higher under Neomycin : 0.924 (p-value = 5.724e-09) for eGFP and 0.882 (p-value = 1.308e-06) for SHBLE, while the same test under Bleomycin displays 0.384 (p-value = 0.094) correlation value for eGFP and 0.414 (p-value = 0.069) for SHBLE.

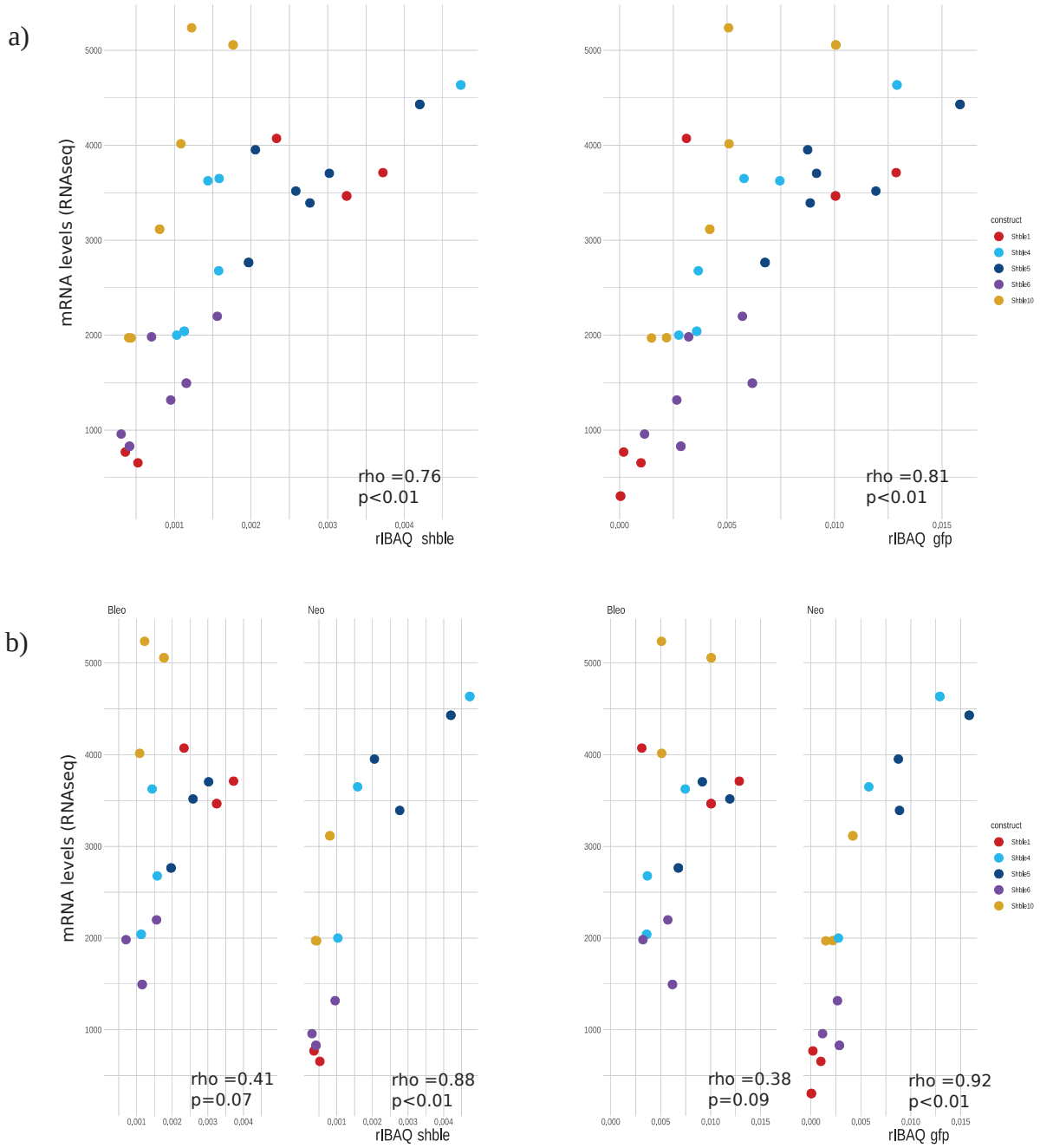
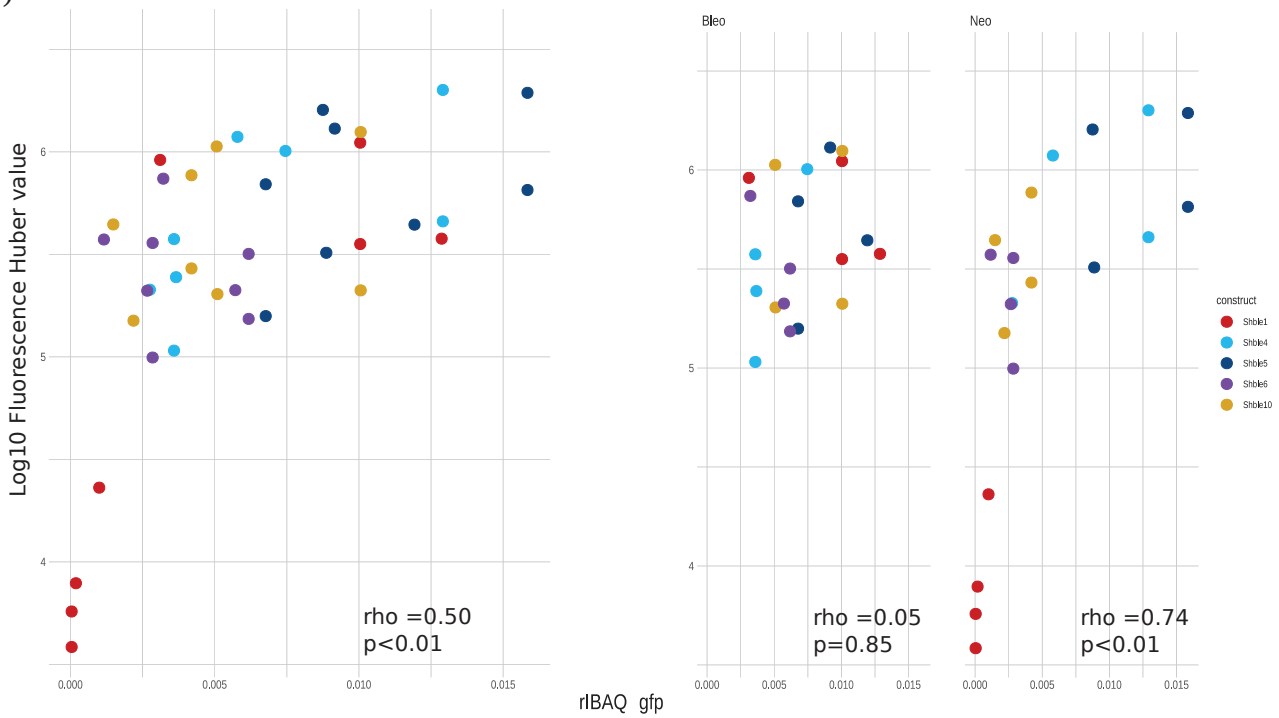


Figure 28: Correlation between mRNA and protein levels *– We calculated the correlation between levels by using Spearman's rank correlation. In each graph results are colored by Construct, and grouped by treatment and by replicate. a) mRNA levels (RNAseq) vs SHBLE and EGFP levels in both treatment b) mRNA levels (RNAseq) vs SHBLE and EGFP levels in separated by treatment*

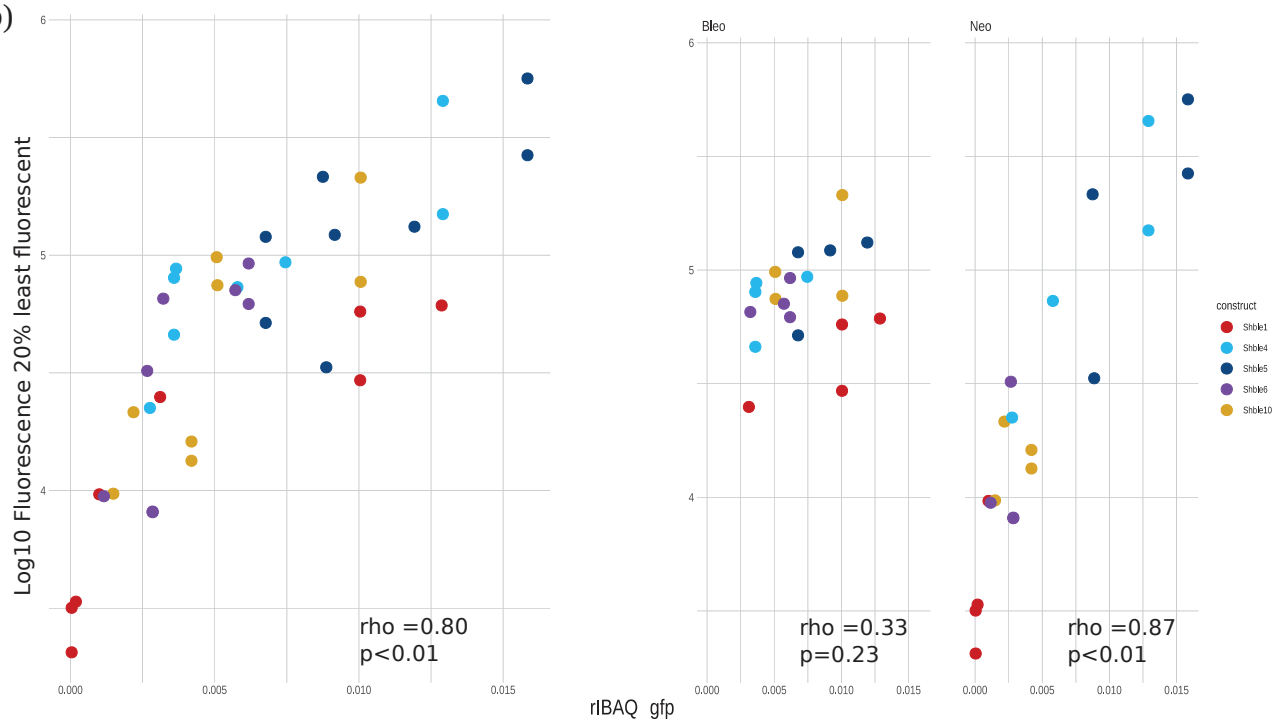
Protein levels *vs* fluorescence intensity levels

Finally we looked at the correlation between variation in EGFP protein levels as estimated by quantitative proteomics and variations in fluorescence intensity as estimated by cytometry (Figure 29). As we saw the emergence of cellular sub-populations in fluorescence levels, we use the three different indicators to describe the fluorescence levels : the median fluorescence value of the 20% most fluorescent cells (upmed), the median fluorescence value of the 20% least fluorescent cells (downmed), and the Huber-M central fluorescence value for the whole population. We used the Huber-M central value for to represent the full population as it is less sensitive to outliers, while the median describes well a smaller portion(20% in this case) of the population. Spearman's rank correlation between the Huber value of the fluorescence of the full population and the protein levels of EGFP is at 0.5 (p-value = 0.005). The same analysis stratified by treatment shows higher correlation values under Neomycin (rho=0.739, p-value = 0.002) and no significant correlation between EGFP levels and fluorescence levels in the Bleomycin treatment. Likewise, the correlation between EGFP protein levels and the fluorescence levels of the 20% least fluorescent cells is only significant in the Neomycin treatment displaying a value of 0.868 under (p-value < 2.2e-16) , and a value of 0.80 rho (p-value = 1.173e-06) when both treatments are considered together. Variations in the fluorescence levels of the 20% most fluorescent cells only correlates with variations in EGFP protein levels under Neomycin (rho = 0.464, p-value = 0.083).

a)



b)



c)

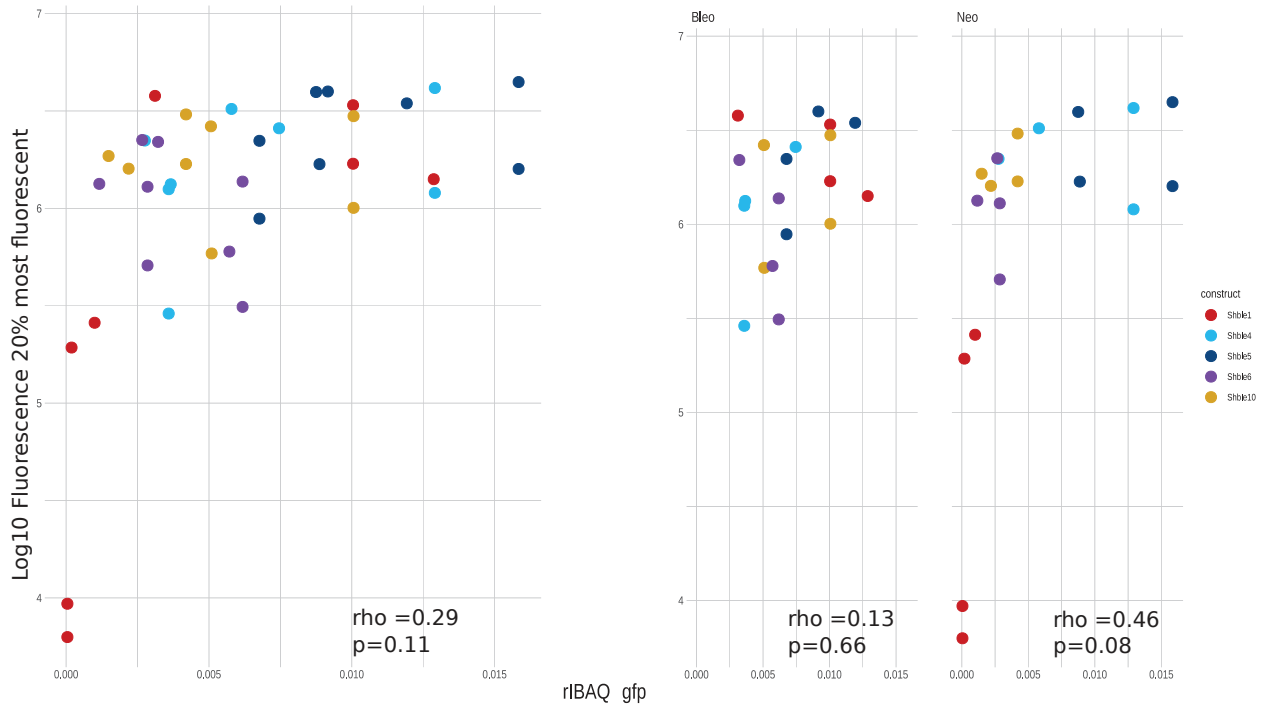


Figure 29: Correlation between the EGFP protein levels and fluorescence levels *– We calculated the correlation using Spearman's rank correlation. In each graph results are colored by Construct, and grouped by treatment and by replicate. The first panel is always the overall correlation, and the second panel is the correlation when considering treatments separately. a) Protein levels (EGFP) vs Huber central value of the fluorescence intensity b) Protein levels (EGFP) vs median of the fluorescence intensity of the 20% least fluorescent cells c) Protein levels (EGFP) vs median of the fluorescence intensity of the 20% most fluorescent cells*

Real-time cell growth measure

In order to estimate the fitness cost of carrying the different versions of the plasmid, and how this changed over time, we performed real-time cell growth measures using an xCelligence system. This allowed us to precisely monitor cell growth under three different conditions : no antibiotics, 400 µg/mL and 2000 µg/mL Bleomycin.

With no antibiotic in the medium cells selected under Bleomycin displayed higher growth rates than under antibiotic, while cells selected with Neomycin have an increasing growth rate at the beginning of the experiment but it drops by the end close to zero just as with 400 or 2000 µg/mL Bleomycin in the media. This is especially true for Shble1, Shble6 and Shble10. Shble4 and Shble5 also show a drop in growth rate, but it is around the same rate as at the beginning of the experiment. Cells selected under Bleomycin have a somewhat better growth rate under 400 or 2000 µg/mL Bleomycin at the end of the experiment than at the beginning(Figure 30).

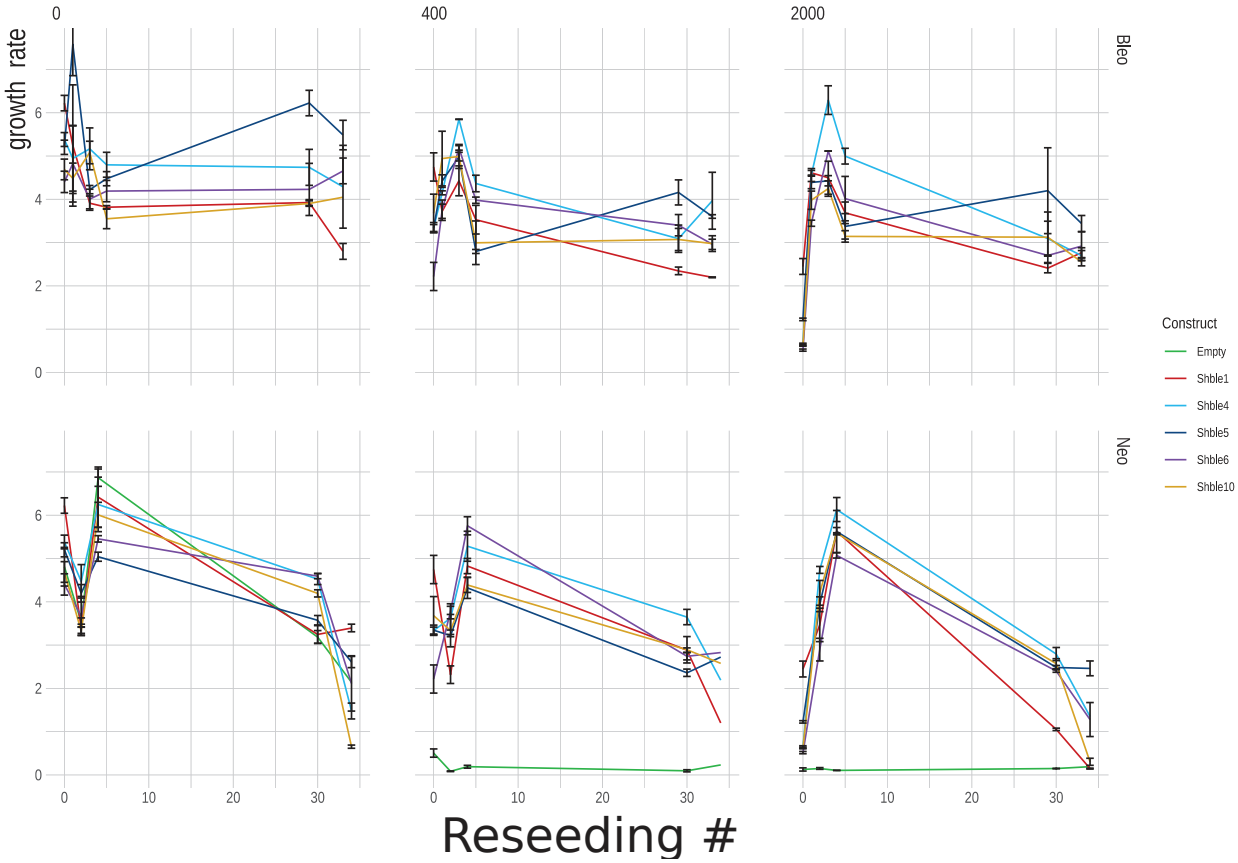


Figure 30: Growth rate variation over time *- growth rate variations generations in varying concentrations of Bleomycin (0, 400, 2000 µg/mL) as measured by xCelligence. Cells are grouped by their original selection treatment (rows).*

Discussion

We conducted a series of multilevel molecular sampling to evaluate the effects of the CUPrefs on the expression of heterologous genes under selection in a long term evolution experiment at different phenotypic levels. We monitored DNA, mRNA, protein and fluorescence levels as well as cell growth kinetics to allow us to quantify whether small synonymous changes on the focal gene impact each step of the gene expression process. We focused on the effect of CUPrefs in the protein synthesis elongation phase, since the modified *shble* genes share the same 5' untranslated regions as well as the coding sequence of the first eight codons.

As our HEK293 cells had evolved for over a hundred generations under three different conditions, we expected to see mutational (or epigenetic), regulatory changes over time, to compensate for the potential cost imposed by the under- or overmatch of CUPrefs in highly expressed heterologous essential genes.

The Bleomycin treatment puts pressure on the cells to express the *shble* gene and thus the *egfp* linked to it, but the rest of the plasmid and the expression of the genes therein encoded are not necessary for cell survival. Additionally the mass production of eGFP might be very disadvantageous as it not only consumes resources, but can also be toxic for cells in high quantities (Ganini et al., 2017). Therefore there is a trade-off, on one hand to efficiently express *shble* even if the CUPrefs do not match tRNA availability, and on the other hand, to not express too much eGFP, even if the CUPrefs of the associated *shble* overmatch tRNA availability and renders therefore the ensemble to be highly expressed. Considering the Neomycin treatment, it imposes a selection pressure to maintain the plasmid to express the NEO_TP enzyme, but not the *shble-egfp* complex. Finally, without antibiotic selection, the selective pressure comes from the weight of carrying and expressing a plasmid and its genes, without any known benefit. In each treatment, there is also a possibility for observing a trans effect, that is the effect caused on other genes present in the cell by the overexpression of the heterologous genes (Frumkin et al., 2018). However, trans-effects of elongation seems to be negligible when cells are not in a stressed condition (amino acid starvation for example) (Firczuk et al., 2013; Racle, Picard, Girbal, Cocaign-Bousquet, & Hatzimanikatis, 2013; Saikia et al., 2016; Shah, Ding, Niemczyk, Kudla, & Plotkin, 2013). In depth analysis of global RNAseq and Proteomics data is still in progress to assess potential competition for Ribosomes or tRNAs in our HEK293 cells. Consequently we will focus here on the cis-effects of heterologous gene expression.

Differences between replicates can be explained by strong bottleneck

R1 and R2 display different results: overall mRNA values are lower, both at the beginning and end of the selection experiment (Figure 21); variation in protein levels doesn't follow the precise same pattern, and above all, fluorescence intensity in terms of sub-populations and changes over time are divergent. In experimental evolution, we should not forget that the driving forces of evolution remain mutation, selection and drift but that the power of selection to act is modulated by population size, history and chance (Lachapelle, Reid, & Colegrave, 2015; Schoustra, Bataillon, Gifford, & Kassen, 2009; Szendro, Franke, De Visser, & Krug, 2013; Weinreich, Watson, & Chao, 2005). Our cells had the same genotypic background as they were unfrozen from the same batch, but the impact of "random" is also dependent of population size (Lachapelle et al., 2015). After each harvesting we reseeded one tenth of the population (approximately $4*10^5$ cells), and thus a strong bottleneck was imposed recurrently on the cell populations every ca. 3.3 generations.. This may explain that although we payed attention to repeat the experiment carefully, the heritable adaptations that might have appeared didn't pass the bottleneck at the same rate in R1 and R2. To explore the potential for parallel evolution for this experimental setup, we are currently repeating it with less constructs and with higher population sizes. Nevertheless we observe comparable trends throughout our replicates, that hint to the same conclusions.

Missing proteins, the limits of Mass Spectrometry

We could detect reads and infer TPM values for 18480 genes, but on the same samples we could only detect peptides to identify 3309 proteins. This disparity may seem surprising at first but it is not unusual, as protein detection is dependent on peptide size and chemical properties that may make it harder to be detected (Ankney, Muneer, & Chen, 2018; Fricker, 2015). Further, by the nature of the technique itself, the next-generation sequencing approach that we applied allows to detect in theory any RNA molecule present in the sample, independently of their sequence. The comparison of the retrieved sequences with the chosen database (the human transcriptome in our case) allows to narrow down the findings by mapping onto a reference, but does not limit the universe of detectable RNA sequences. On the contrary, in the case of the unlabeled quantitative proteomics, the peptides detected are not sequenced. Instead, they are characterized by a mass and a charge/mass ratio, and this information is contrasted against the universe possible peptides generated after hydrolysis from a chosen protein database (the human proteome in our case). This means that only peptides with sequences known in forehand are detectable, and that any chemical modification leading to changes in the mass or charge/mass ratio may render a peptide, and *in fine* a protein, undetectable. In our hand, the most conspicuous case of the lack of match between

transcriptomic and proteomic data refers to the NEO_TP protein: although our cells survived under Neomycin selection for seven months, there was no NEO_TP protein detected by the mass spectrometry measures despite its mRNA being detected in RNAseq (Figure 31). It is unclear, if this enzyme is especially prone to degradation or it could not be detected because of its peptide's properties, the latter being more plausible. Furthermore, we observed a constant 1:3 SHBLE:EGFP ratio at the protein level, as opposed to an expected 1:1 ratio as in theory EGFP can only be translated after SHBLE. It can be proposed that Ribosomes could skip the *shble* part of the mRNA and start directly at *egfp*, in a behavior known as leaky scanning (Ryabova, Pooggin, & Hohn, 2006). This is unlikely, first because the different CUPrefs should for this have an effect on the ability of the corresponding mRNA to engage Ribosomes and initiate translation sequence, but all constructs share the 5'UTR and the first eight codons; and second because the 1:3 ratio is maintained throughout constructs and selection regimes. It is more probable that SHBLE is more difficult to detect by mass spectrometry than EGFP, resulting in lower rIBAQ values across samples.

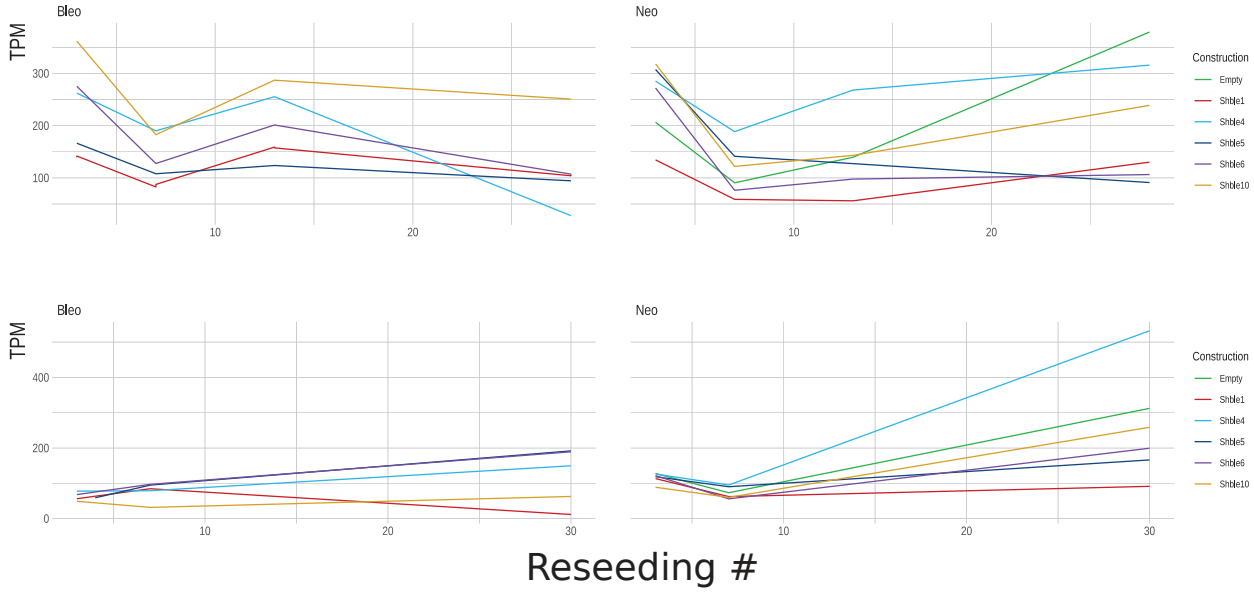


Figure 31: NeoR mRNA levels over time - RNAseq *– NeoR RNAseq results of R1 and R2 for each construct and each treatment over time. In y-axis the reseeding number i.e. sampling time-points, in x-axis the TPM.*

Transfected cells reach and maintain high resistance to Bleomycin, without apparent fitness cost

In the real-time growth experiment, we measured the growth rate of the cell lines in presence of Bleomycin at several stages of the Selection experiment. Our results show that cells evolved under Neomycin grow slower at the end of the experiment even without antibiotics in the media, while the ones evolved under Bleomycin display the same growth rate after ca. ten and 100 generations, without any notable difference between 400 or 2000 µg/mL Bleomycin in the media. Considering the absence of antibiotics, the difference in maximum growth is notable, especially because we did not observe longer intervals between reseedings of cells from the Neo treatment than in the Bleo treatment. As the xCelligence measures run for 72h but cells are reseeded every 7 days on average, it is conceivable that their growth is initially slower and could speed up after a few days. However we cannot conclude on the fitness cost of the plasmids with different CUPrefs.

Transcription of the spliced forms of Shble4 and Shble6 is low

In the Day2 experiment it was observed that spliced forms of Shble4 and Shble6 made up respectively 35% and 80% of all *shble-egfp* mRNAs. In the selection experiment however, in both treatment from the very beginning we recovered very low transcript levels for these spliced forms (mean of overall measures with respect to the total *shble-egfp* transcripts: Shble4 =4%,Shble6 = 30%). This suggests that the expression of the spliced forms is quite rapidly limited by the cells in both treatment, more precisely after three reseedings (ca. 10 generations) these low levels are already generalized in all samples independently of the selection regime. Proteins from the spliced form are also absent in the samples, nevertheless it must be considered that the potential SHBLE* spliced forms differed only by one single possible peptide from the full-length SHBLE, thus rendering detection more complicated. Curiously we see a drastic increase of mRNA spliced forms in R2 Neo Shble6, accompanied by a decrease of the non-spliced form. This event seems to be unique in the data. Sequencing a close time-point or setting up a splice variant-specific rt-qPCR could reveal if it is a technical error, or a real increase. That said, as under Neomycin the SHBLE protein is not necessary for survival, raising the number of spliced forms, could have been a winning event if it allows to down-regulate its expression and limit the metabolic cost of keeping the full plasmid and of expressing heterologous genes.

Differences in CUPrefs do not explain variation in transcription and translation levels under Neomycin

In the Neomycin treatment, where expression of the *shble-egfp* complex is not necessary for survival, we observe different phenotypes specific to each construct. They can be categorized into three groups : Loss of gene expression (Shble1), Low gene expression (Shble6, Shble10), High gene expression (Shble4, Shble5, Empty). These categories can be observed throughout measures as correlations between each level are high, with DNA-RNA being the lowest especially in R2 (Spearman's rank correlation rho : R1 = 0.7, $p<0.01$; R2 = 0.49, p<|0,01). Although variation in DNA levels explains partially variation in mRNA and thus in Protein and Fluorescence, as the same categories have been recovered in the two replicates, it is unlikely that the phenotypes are merely the result of inequality in transfection at the launch,especially because transfection efficiency was optimized for in pilot experiments.

Curiously, when compared with the transcript levels for the *neo_tp* mRNA, the former described categories are maintained with the exception Shble5 displaying a very low TPM, close to Shble1 and Shble6. Unfortunately, as NEO_TP was not detected by mass-spectrometry, we cannot quantify translation levels for this protein. This is quite counter-intuitive, as in theory there could be selection for increasing the expression of the plasmid, leading to high *neo_tp* TPM levels to resist Neomycin, and as a side product high SHBLE and EGFP levels by genetic hitchhiking. But the low transcript levels of *neo_tp* in Shble5, opposed to it's high EGFP and SHBLE levels do not support this interpretation.

In any case these categories based on mRNA, protein and fluorescence levels are not in direct correlation with GC3 content or CUPrefs (Figure 32). Most likely the complexity of eukaryotic gene expression and all its steps obscures the effects of CUPrefs, in conditions in which the modified heterologous genes are not under direct strong selective pressure.

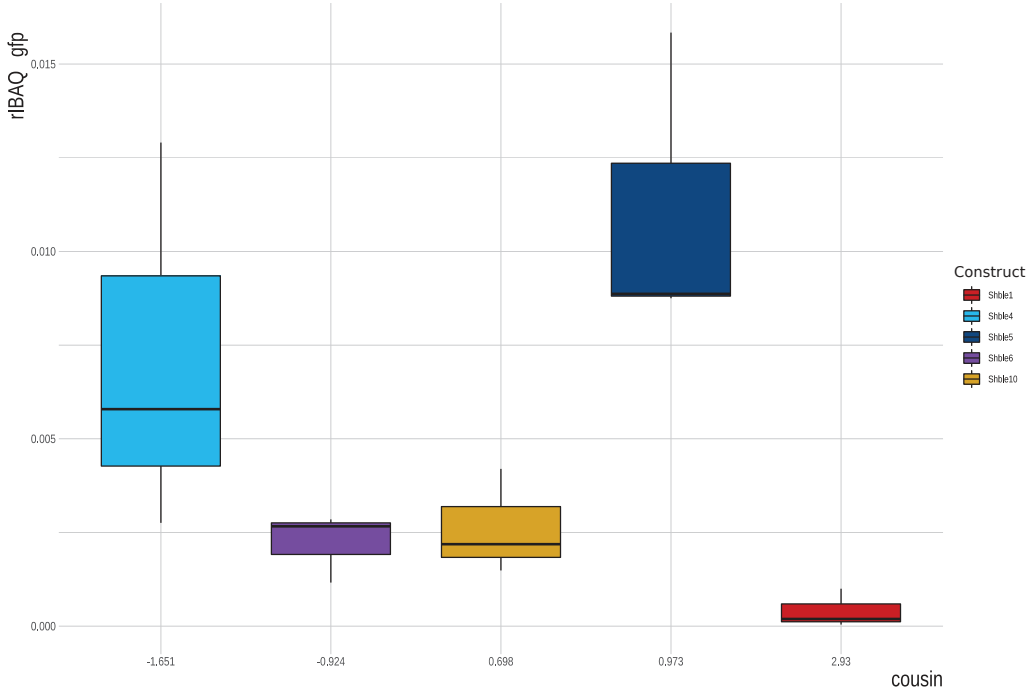


Figure 32: Overall protein levels vs COUSIN score *– rIBAQ of EGFP in the different constructs ranked by their COUSIN score.*

Heavy selective pressure leads to convergent phenotypes under the Bleomycin treatment

When considering fluorescence, protein levels, and mRNA levels, we observe a recurrent phenomenon: cells transfected with different constructs behave differently from each other under Neomycin, while under Bleomycin they display convergent behavior. This phenotypic convergence between cells carrying different constructs is observed throughout different molecular integration levels, and for the two replicates, even if most of our measures reflect average values of the corresponding variable for the full population, and only fluorescence conveys results for each individual cell in a sample. Our results show that, over time even cells carrying constructs with undermatching CUPrefs to those in the average human genome reached fluorescence levels comparable to those in cells carrying over-matching ones. This suggests that in this setup the fitness cost of high expression, competition for tRNAs, and the potential toxicity of eGFP are negligible compared to the advantage of efficiently expressing *shble*. This layout evokes a scenario similar to that of essential genes in bacteria, that are under negative purifying selection and thus more conserved (Dilucca, Cimini, & Giansanti, 2018; Jordan, Rogozin, Wolf, & Koonin, 2002). In our case it seems that cells reached a high level of SHBLE production needed to counter the effect of Bleomycin, and any change that would lower EGFP and with it SHBLE is not viable or could not be fixed yet. A longer time scale, higher population sizes or a more permissive bottleneck could probably allow for the apparition of additional phenotypes. Although our HEK293 cells are not haploid, their multiplication and replication is asexual. The probability of fixing a beneficial mutation in an asexual population is a decreasing function of both population size and mutation rate (Gerrish & Lenski, 1998). As cells in this experiment have most likely multiple copies of the plasmids and thus heterologous genes, we can assume that it buffers the effect of deleterious mutations, but also lowers the chance of the substitution of the population by variants with beneficial mutations.

Still, we do note a stable cellular sub-population with low fluorescence in the Shble1 lineages, but this secondary population does not increase, nor takes over the whole population which would be the case if it carried an advantageous adaption with an important selective coefficient (consideration made also for the impact of drift imposed by our recurrent bottlenecks). Moreover, the origins of these sub-populations are not clear to us yet. We hypothesize that it may be a side effect of the occasional high confluence of cells at the moment of harvesting. We are currently testing if the density of cells at the time of harvesting are not at the cause, to exclude a potential effect of manipulation.

Low correlation between genome, transcriptome and proteome, hints at both pre- and post-transcriptional regulations under Bleomycin

We assessed the correlation between the measures of each step of gene expression, and found that it is systematically lower in the Bleomycin treatment than under Neomycin. In fact the variation in DNA levels explains 45% of the variation in mRNA levels of heterologous genes in R1 and mRNA levels explain 40 % of the variation in protein levels. This clearly suggest that other factors beyond plasmid copy number influence transcription and translation. Considering that each step is characterized by low correlation values, we propose that both pre- and posttranscriptional mechanisms intervene to achieve sufficient SHBLE expression. We hope to uncover some of these mechanisms by further examining the correspondence between the full transcriptome and proteome of the cells, not focusing on the heterologous genes alone this time, but on elements playing a role in mRNA and protein degradation, transcription and translation elongation, and other regulatory mechanisms.

Cells carrying Shble1 lose *egfp* expression

The Shble1 synonymous version used is the one with the highest GC3 value (93%), and the highest COUSIN value (2.93), meaning that it largely over-matches the human CUPrefs. These characteristics make it the construct with the highest SHBLE initial expression. The high expression of both SHBLE and EGFP from cells carrying the Shble1 version may use up resources and limit Ribosomes for other genes. Indeed very early in R1 and by the end of R2, we see Shble1's fluorescence decrease to levels comparable to that of the non-fluorescent "Mock" cell line. Along its fluorescence intensity, quantified eGFP protein and mRNA levels also drop, or stay low in the case of rt-qPCR results. Moreover, despite the DNA still seems to be present (see qPCR), the RNAseq data show a loss over time of the eGFP part of the *shble-egfp* complex. Although we could not evaluate the fitness cost of carrying the different version of the plasmid, Shble1 losing fluorescence in both replicates seems to confirm that over-matching CUPrefs and the expression patterns linked to it are under selective pressure. Based on these results we have launched a smaller set of experiments with five clonal populations transfected with the Shble1 Construct and maintained under Neomycin. Under these conditions, these cell lines repeatedly lose fluorescence, demonstrating that this phenomenon is reproducible. We propose that regulatory mechanisms, involving or not mutations in the cellular genome, inhibit the *egfp* transcription or that posttranscriptional modifications induce partial cleaving or degradation of the mRNA reducing the overexpression metabolic burden as well as the potential trans-effect of the high expression of

unprofitable genes. Another possibility could be that the plasmid might have integrated in the cellular genome immediately downstream the shble open reading frame, losing a part of its backbone and leading to a loss of the egfp moiety of the mRNA. We aim to further explore the status of the plasmid DNA by means of flanking sequencing.

Conclusion

In this study we aimed at exploring the effects of CUPref of heterologous genes in eukaryotic cells, and how the cells may compensate for the differential burden imposed by the expression of these different synonymous genes over 100 generations. Despite the challenges of long-term experimental evolution and the complexity of our multilevel dataset we could identify some impact of CUPrefs on the cells. We show, that if the expression of the modified genes is directly under selection, cells overcome without any notable cost the CUPrefs mismatch, and in spite of the differences, converge to similar expression patterns. On the other hand, when the modified genes are subject to genetic hitchhiking, potential regulatory mechanisms create different expression profiles to limit metabolic cost, to the point of completely silencing the translation and probably transcription of the gene in question.

Perspectives

As mentioned above, several analysis and control experiments are still in progress. Among others with the complete RNAseq and proteomics data at hand, Come Morel and Arthur Jallet have started to explore the potential compensatory mechanisms and trans-effects of high heterologous gene expression. We have started to look into regulatory pathways that have been activated, and also at variations in tARN synthetases levels. We also intend to evaluate protein levels of chaperons involved in reparations of misfolded proteins. Another effect we plan to study more in detail, is the competition for resources, and the effect of expressing heterologous genes on other genes that have similar CUPrefs to the heterologous genes, or might be out competed for tRNAs or Ribosomes. For this, in addition to the data we already collected, it would be interesting to test a similar setup, but with poor cell media, to induce stress and starvation in the cells, as it was shown that CUPrefs have a significant effect on translation elongation during amino acid starvation (Saikia et al., 2016).

In parallel with the selection experiment, a competition experiment was also performed to properly evaluate cellular fitness under the different conditions and for the different constructs. In this experiment we placed the created cell lines in competition with each other under Neo and Bleo

treatment. The cells were under selection of Bleo and Neo for ten reseedings, then mixed together. Cells were harvested the same way as for the selection experiment, with the intent to sequence DNA and mRNA to quantify the ratio of different cell lines in a population. This experiment will give us a more complete picture of the fitness cost of and benefits of the different Constructs.

Finally, as new technologies emerges, tRNA sequencing could give more insight into how the cells compensated for the non-matching CUPrefs, and it could complete our experiment with one more stage of translation (Smith, Abu-Shumays, Akeson, & Bernick, 2015).

Overall, the next step in the project, is to analyze and exploit the transcriptomic and proteomic data already available, and to repeat and refine some aspects to ensure reproducible results

Chapter 5

General Discussion

With the raising concern of climate change, globalization and humanity's impact on nature, there is an increasing risk of the emergence of pathogens. Indeed, as natural environments are transformed into fields or residential areas, contact between humans and wildlife and their pathogens, are on a rise. Meanwhile mono-cultures and intensive livestock industry create a perfect terrain for diseases to develop, as genetic diversity is low, and population density is high. Together these factors augment the chances of a spillover event and the emergence of zoonotic diseases (Bengis et al., 2004; Johnson et al., 2020; Jones et al., 2013; Rohr et al., 2019).

In this context, to better understand how pathogens, and more precisely viruses make use of the host's cell machinery, is essential. In the recent pandemic caused by the SARS-CoV2 a great effort has gone into the complete dissection of its genome and functioning. Among other works, the study of the evolution of SARS-CoV2 CUPrefs offers us insight into its natural history and a glance on its potential co-evolution with its new host : *Homo sapiens*. Compared to other human infecting *Coronaviridae*, the three new strains involved in the COVID-19 outbreak aren't close to their natural mutational equilibrium, however they are becoming AU richer as a preponderance of C→U mutation has been shown by Simmonds (Simmonds, 2020), potentially to reach a new equilibrium (Daron & Bravo, 2021). In this case the contribution of CUPrefs to the zoonotic nature of coronaviruses seems to be minor. Other authors described a trans-effect on the host, as SARS-CoV2 seems to down-regulate the expression of host genes with similar CUPrefs in CACO-2 cells (Alonso & Diambra, 2020; Bojkova et al., 2020).

In other viral families, the role of CUPrefs suggests correlation with other primary viral traits, such as infection phenotype. Human Papillomaviruses (HPVs), are a global health concern, as certain oncogenic HPVs are linked to the development of cervical cancer (Forman et al., 2012). Upon analyzing their CUPrefs, Félez-Sánchez and coworkers found that most HPVs genes do not match human CUPrefs, except the genes encoding for capsid proteins in viral genotypes that are linked to productive lesions (Félez-Sánchez et al., 2015). Furthermore, in this study the factor contributing the most to explain variation in CUPrefs of HPVs, was the phenotype of the viral infection. In our study on Cetartiodactyla PVs, we do not observe a direct link between the clinical traits and CUPrefs of the virus, instead we found a correlation with gene expression pattern, inline with other works on HPVs and other viruses (B. Miller, Hippen, M. Wright, Morris, & G. Ridge, 2017). These examples further confirm, that CUPrefs, play an important role in viral evolution and in host-virus interaction via the expressed protein, and it is indispensable to fully comprehend viruses.

Growing availability of OMICS data online and the fast development of bioinformatic tools offers an excellent opportunity to compare and study the effect of codon changes in various organisms and settings. To cite a few : E. Lara-Ramìrez and co-workers analyzed 3047 Dengue virus (DENV) sequences and CUPrefs in their study (Lara-Ramírez et al., 2014). They revealed that mutational bias and purifying selection are the main forces driving the codon usage in DENV, but with distinct pressure on specific nucleotide position in the codon. In an other work, Mordstein and her team, inspected 1520 vertebrate infecting viruses, and the factors that shape their CUPrefs (Mordstein et al., 2021). They propose, that CUPrefs are under the influence of several factors, for example: the nature of the genetic material, location of the viral replication in the host cell and immune-evasion. In our work we also made use of the advantages of the large and open databases of omics data although in a somewhat smaller scale : in the second chapter we analyzed in depth the sequences of 58 Papillomaviruses infecting Cetartiodactyles, and in the third chapter we explored the evolving CUPrefs of the Polypyrimidine Tract Binding Protein and its paralogs, in a selected pool of 74 vertebrates.

On the other side of the coin, are the organisms that harbor viruses, in our case, eukaryotic hosts. Eukaryotic organisms enclose an intricate and delicate pipeline for gene expression that is equipped with multiple regulatory mechanisms and fine tunings. One of these, are the frequency and choice of synonymous codons in the DNA sequence. In prokaryotes, transcription and translation occur in the same place, and often simultaneously (McGary & Nudler, 2013). This means that the mRNA goes through much less maturation compared to what we see in eukaryotes, therefore the effect of CUPrefs in gene expression might be more evident in prokaryotes. Meanwhile the study of CUPrefs in eukaryotes has to consider the many mechanisms a cell wields to regulate expression from the very beginning of transcription initiation, till the degradation of proteins. In this work we focused on eukaryotic CUPrefs evolution over generations, and how do they react to heterologous gene CUPref variations, in order to be able to better interpret their relationship with infecting viruses.

To do so we looked at paralogous gene evolution in vertebrates and carried out a long-term selection experiment using human cells (HEK293) transfected with synonymous versions of an antibiotic resistant gene. Although there is some debate over which forces shape CUPrefs in vertebrates, in our study we found that depending on the genes function and expression pattern in time and space, several factors shape CUPrefs in higher eukaryotes and it cannot be explained by just mutational bias or just translational selection. Local mutation bias, GC-biased gene conversion and translational selection both act in the case of paralogous genes over several million years, endowing the paralogs with distinct CUPrefs, as shown with the example of PTBPs. However our experimental evolution setup shows, that on a smaller time scale, these forces are negligible, but

epigenetic regulatory processes are quick to adjust expression patterns, if needed, despite and independently of CUPrefs. When mutation and translational selection do have the time and the power to act and a large enough population, in both viruses and vertebrates, they may shape a gene's CUPrefs to fit to its expression pattern and function. For example, in PTBPs there are signs for tissue specificity, and each paralog's CUPrefs seem to follow the CUPrefs of its environment (see chapter 3). In Papillomaviruses we observe divergent CUPrefs between genes that are expressed in the early and the late stage of infection (See chapter 2). In the same PVs we observed correlation between the presence of conserved regulatory motifs, and host species. This may hint to co-evolution between host, and pathogen transcription and translation initiation, as eukaryotes seem to intervene efficiently in the expression of heterologous genes. By contrast, if the expression (or non-expression) of a heterologous gene is not an immediate threat, or too costly metabolically, with a strong promoter and the right regulatory motifs, it may still be expressed (see chapter 4).

Viral CUPrefs of human-infecting viruses usually go against host codon bias, and it was proposed that it is in part to avoid the host immune system (Mordstein et al., 2021). In our Selection experiment, the cells did not had an adaptive immune system, but in the case of Shble1, the cells partially silenced transcription of the over-matching *shble-egfp* complex. We propose that, even in the absence of an immune response, cells may epigenetically down-regulate viral genes if they harbor an immediate high metabolic cost. Meanwhile viral genes with undermatching CUPrefs may be expressed in the background without further alarming the immune system or the regulatory mechanisms of the host cell. Of course codon usage preferences are but small mechanisms in the insanely complex machine of virus-host interactions, still, we cannot comprehend the full picture without them.

Conclusion

In this work, we seek to enlarge our knowledge of the role of CUPrefs in viral and eukaryotic gene expression and evolution, and how they interconnect. We offer a multi-faced study, where we dual-wield *in vivo* and *in silico* analysis, completed with experimental data collected from each step of gene expression. Overall we found that the CUPrefs play a role in regulating expression in terms of its differed time or place, as seen in both Cetartiodactyla PVs and vertebrates. Meanwhile, we show that Eukaryote cells can adjust rapidly by complex regulatory mechanisms to overcome the burden imposed by the overexpression of of heterologous CUPrefs if they are needed for survival, or down-regulate them if their expression is costly. However, in a long term selection experiment we show that a with a strong promoter genes with undermatching CUPrefs can be maintained and expressed over a 100 generations, even if it offers no positive effect on the fitness.

Chapter 6 – Additional work

Loop 1 of APOBEC3C Regulates its Antiviral Activity against HIV-1

The work is part of a long-term collaboration of our team and the team of Carsten Münk at the University Düsseldorf on the evolution, diversity and function of APOBEC3 genes in primates. In this specific case, our collaboration consisted in providing the phylogenetic analyses and ancestral state reconstruction. The article was published in the *Journal of Molecular Biology* in 2020.

The APOBEC3 family (A3) consists of single stranded (ss) DNA cytidine deaminases, that assure immune defense against retroviruses, retrotransposons and other viral pathogens . As a results of several duplication during primate evolution, today we can find various versions of this protein, that has either one or two zinc-coordinating DNA cytosine deaminase motif.

A3C has been shown to inhibit viral particles by incapsidating into them and deanimate Cytidines into Uridines during retros-transcription. Human A3C (hA3C) for example is known to inhibit *Simian immunodeficiency virus* infecting the African green monkey and the Rhesus macaque. Curiously, against *Human immunodeficiency virus 1* , A3C has only shown a limited restrictive capacity and findings on the subject are often contradictory.

In this study, a synthetic A3C-like protein that has a high restraining capacity against HIV-1 was created, and researchers collaborated to uncover its underlying mechanisms and characteristics. The synthetic A3C-like is a hybrid of smmA3C and smmA3F from the Sooty mangabey monkey, the two sequences are highly similar.

In order to test if other non-human primates can resist HIV-1, its viral infection factor (*vif*) was marked with luciferase, and cells with different versions of A3 (human, non-human primates and synthetic A3C-like) were infected with it the virus. Infection was quantified two days later, and has shown that the synthetic A3C-like protein restricts 10 folds the HIV-1 infection compared to the human and other monkey A3Cs. To identify the regulatory domain that meditates restriction, chimera sequences were created from hA3C ans smmA3C- like. The chimera with the most similar activity to smmA3C like was C2, which is a hA3C sequence, with a swap of 36 residues at the N' terminus. N-terminal motifs were mutated, and our collaborators pinpointed the RKYG motif as controller of the antiviral activity. In hA3C it is the WE-RK mutation in loop 1 that enhances interaction with ssDNA and thus offering a strong deaminase dependent antiviral function. Unexpectedly it was shown by experimental results that A3C-WE-RK expression also strongly inhibits human LINE-1 retrotransposition activity.

Phylogenetic analysis of A3Z2 loop 1 regions in primates suggests two duplication events that occurred after the divergence between New World Monkeys, and Old World Monkeys 43.2 Mya, but before the diversification between Old World Monkeys and Apes.

We proposed that the two series of ancestral gene duplications that generated A3C, A3D-CTD and A3F-CTD allowed neo/subfunctionalisation: A3F-CTD maintained the ancestral RK residues in loop 1, while diversifying selection resulted in the RK-WE modification in Old World anthropoid's A3C, possibly allowing for novel substrate specificity and function.

Research Article

Loop 1 of APOBEC3C Regulates its Antiviral Activity against HIV-1

Ananda Ayyappan Jaguva Vasudevan [1*†], Kannan Balakrishnan [1,2†], Christoph G. W. Gertzen [3,4,5], Fanni Borvető [6], Zeli Zhang [1‡], Anucha Sangwiman [1], Ulrike Held [7], Caroline Küstermann [7], Sharmistha Banerjee [2], Gerald G. Schumann [7], Dieter Häussinger [1], Ignacio G. Bravo [6], Holger Gohlke [3,4] and Carsten Münk [1*]

*1 - **Clinic for Gastroenterology,** Hepatology, and Infectiology, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany*
*2 - **Department of Biochemistry,** School of Life Sciences, University of Hyderabad, Gachibowli, Hyderabad, India*
*3 - **Institute for Pharmaceutical and Medicinal Chemistry,** Heinrich Heine University Düsseldorf, Düsseldorf, Germany*
*4 - **John von Neumann Institute for Computing (NIC),** Jülich Supercomputing Centre & Institute of Biological Information Processing (IBI-7: Structural Biochemistry), Forschungszentrum Jülich GmbH, Jülich, Germany*
*5 - **Center for Structural Studies (CSS),** Heinrich Heine University Düsseldorf, Düsseldorf, Germany*
*6 - **Centre National de la Recherche Scientifique,** Laboratory MIVEGEC (CNRS, IRD, Uni Montpellier), Montpellier, France*
*7 - **Division of Medical Biotechnology,** Paul-Ehrlich-Institute, Langen, Germany*

Correspondence to Ananda Ayyappan Jaguva Vasudevan and Carsten Münk: Structural Cell Biology Group, Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), NIH, Research Triangle Park, NC 27709, USA (A.A. Jaguva Vasudevan). Clinic for Gastroenterology, Hepatology, and Infectiology, Medical Faculty, Heinrich Heine University Düsseldorf, Building 23.12.U1.82, Moorenstr. 5, 40225 Düsseldorf, Germany (C. Münk). *anand.jaguvavasudevan@nih.gov (A.A. Jaguva Vasudevan), carsten.muenk@med.uni-duesseldorf.de (C. Münk)*
https://doi.org/10.1016/j.jmb.2020.10.014
Edited by Eric O. Freed

Abstract

APOBEC3 deaminases (A3s) provide mammals with an anti-retroviral barrier by catalyzing dC-to-dU deamination on viral ssDNA. Within primates, A3s have undergone a complex evolution *via* gene duplications, fusions, arms race, and selection. Human APOBEC3C (hA3C) efficiently restricts the replication of viral infectivity factor (*vif*)-deficient *Simian immunodeficiency virus* (SIVΔ*vif*), but for unknown reasons, it inhibits HIV-1Δ*vif* only weakly. In catarrhines (Old World monkeys and apes), the A3C loop 1 displays the conserved amino acid pair WE, while the corresponding consensus sequence in A3F and A3D is the largely divergent pair RK, which is also the inferred ancestral sequence for the last common ancestor of A3C and of the C-terminal domains of A3D and A3F in primates. Here, we report that modifying the WE residues in hA3C loop 1 to RK leads to stronger interactions with substrate ssDNA, facilitating catalytic function, which results in a drastic increase in both deamination activity and in the ability to restrict HIV-1 and LINE-1 replication. Conversely, the modification hA3F_WE resulted only in a marginal decrease in HIV-1Δ*vif* inhibition. We propose that the two series of ancestral gene duplications that generated A3C, A3D-CTD and A3F-CTD allowed neo/subfunctionalization: A3F-CTD maintained the ancestral RK residues in loop 1, while diversifying selection resulted in the RK → WE modification in Old World anthropoids' A3C, possibly allowing for novel substrate specificity and function.

© 2020 Elsevier Ltd. All rights reserved.

Introduction

The APOBEC3 (A3) family of single-stranded (ss) DNA cytidine deaminases builds an intrinsic immune defense against retroviruses, retrotransposons, and other viral pathogens[1–4] There are seven human A3 proteins that possess either one (A3A, A3C, and A3H) or two (A3B, A3D, A3F, and A3G) zinc (Z)-coordinating DNA cytosine deaminase motifs. Z motifs can be classified into three groups (Z1, Z2, Z3), but share the consensus signature HXE[$X_{23\text{-}28}$]PC[$X_{2\text{-}4}$]C (where X indicates a non-conserved position).[5–9] A3C is the only single-domain A3Z2 protein in humans. During primate evolution, the ancestor of the *A3C* gene duplicated several times and formed double-domain *A3Z2-A3Z2* genes, which are *A3D* and *A3F*.[6] Initially, A3G was characterized as the factor capable of restricting infection of HIV-1 lacking *Vif* (viral infectivity factor) protein in non-permissive T cell lines and its biochemical properties and biological functions have been extensively studied.[3,10–13]

The encapsidation of A3s into the viral particles is crucial for virus inhibition.[14–19] During reverse transcription, viral core-associated A3 enzymes can deaminate cytidines (dC) on the retroviral ssDNA into uridines (dU). These base modifications in the minus-strand DNA cause coding changes and premature stop codons in the plus-strand viral genome (dG → dA hypermutation), which impair or suppress viral infectivity.[2,11,20–23] In addition to the mutagenic activity of the virus-incorporated A3s, deaminase-independent mechanisms of restriction have been identified such as impeding reverse transcription or inhibiting DNA integration.[24–29] To counteract A3 mediated inhibition, lentiviruses evolved the Vif protein, which physically interacts with A3s, targeting them for polyubiquitination and proteasomal degradation.[30–32] These A3-Vif interactions are often species-specific and an important factor reducing virus cross-species transmission.[33–38]

In addition to A3G, A3D, A3F, and A3H were shown to restrict HIV-1 lacking *vif* (HIV-1Δ*vif*).[2,37,39–42] Recently, mutation signatures resulting from the catalytic activity of nuclearly localized A3s (especially A3A, A3B, and likely A3H) were reported in several cancer types.[43–50] The knowledge about A3C is rather sparse. A3C is distributed in both cytoplasm and nucleus[51] and does not seem to be a causative agent of chromosomal DNA mutations. In addition, human A3C is known to act as a potent inhibitor of *Simian immunodeficiency virus* from African green monkey (SIVagm) and from rhesus macaque (SIVmac), limits the infectivity of herpes simplex virus, certain human papillomaviruses, murine leukemia virus, *Bet*-deficient foamy virus, and hepatitis B virus, and represses the replication of LINE-1 (L1) endogenous retrotransposons.[51–61] In contrast, the restrictive role of A3C on HIV-1 is marginal and there are several contradictory findings regarding its viral packaging and cytidine deamination activity.[42,52,62–64] Notably, A3C is ubiquitously expressed in lymphoid cells,[5,52,65,66] mRNA expression levels of A3C are higher in HIV-infected $CD4^+$ T lymphocytes;[42,52] and significantly elevated in elite controllers compared to ART-suppressed individuals.[67] A3C was found to moderately deaminate HIV-1 DNA if expressed in target-cells of the virus with the effect of increasing viral diversity rather than causing restriction.[65]

The crystal structure of A3C and its HIV-1 Vif-binding interface has been solved.[68] The study revealed several key residues in the hydrophobic V-shaped groove formed by the α2 and α3 helices of A3C that facilitate Vif binding, resulting in proteasome-mediated degradation of A3C.[68] We have extended this finding and identified additional Vif interaction sites in the α4 helix of A3C.[69] Apart from a previous study that predicted putative DNA substrate binding pockets;[57] biochemical and structural aspects of A3C enzymatic activity and their relevance for antiviral activity remain hitherto not well investigated.[3,4]

Recently, we have shown that increasing the catalytic activity of A3C by an S61P substitution in loop 3 is not sufficient to restrict HIV-1Δ*vif*.[70] It is unknown why A3C can potently restrict SIVΔ*vif* while HIV-1Δ*vif* is largely resistant, despite the fact that wild-type (WT) human A3C possesses reasonable catalytic activity and is encapsidated efficiently into retroviral particles.[70] Here we set out to further explore the determinants of A3C's restrictive capacity of HIV-1. We generated a synthetic open reading frame derived from sooty mangabey monkey genome (smm, *Cercocebus atys (torquatus) lunulatus*) coding for an A3C-like protein (hereafter called smmA3C-like protein) capable of restricting HIV-1 to similar or higher extents than human A3G. This A3C-like protein was reported to be resistant to HIV-1 Vif-mediated depletion.[69] Using this smmA3C-like protein as a tool, we dissected a novel structure–function relationship of hA3C and discovered the importance of loop 1 for A3C to achieve strong inhibition of HIV-1.

Results

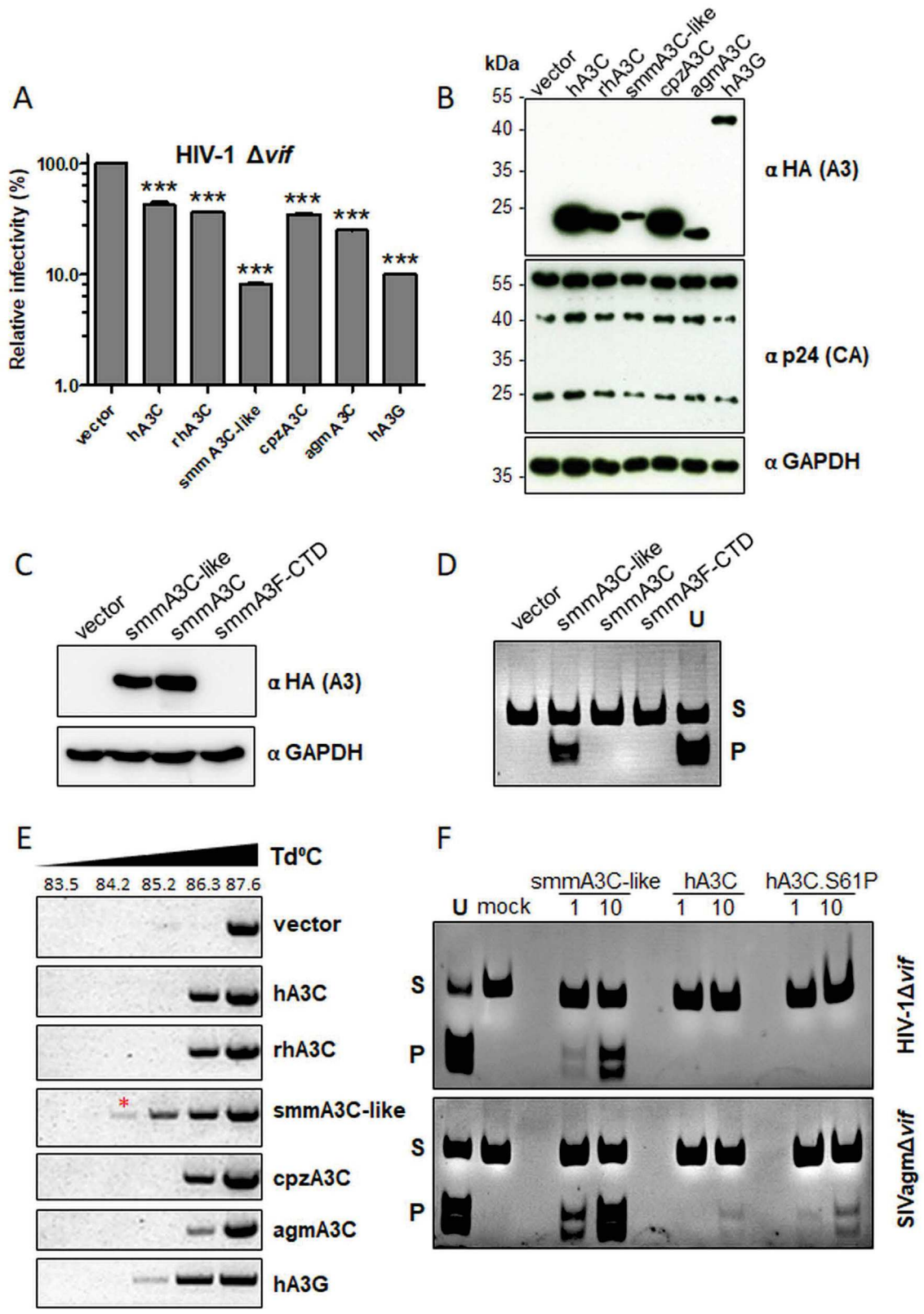
Identification of an A3Z2 protein with enhanced antiviral activity

To determine whether A3C from non-human primates can potently restrict HIV-1Δ*vif* propagation, we produced HIV-1Δ*vif* luciferase reporter virus particles with A3C (an A3Z2 protein) from human, rhesus macaque, chimpanzee (cpz), African green monkey (agm), and with human A3G (an A3Z2-A3Z1 double domain protein), or with a synthetic smmA3C-like protein and tested the infectivity of the respective viral particles. Viral particles were pseudotyped with the glycoprotein of *Vesicular stomatitis virus*

(VSV-G) and normalized by reverse transcriptase (RT) activity before infection. The firefly luciferase enzyme activity of infected cells was quantified two days post infection. Figure 1(a) shows the level of relative infectivity of HIV-1Δ*vif* in the presence of the tested A3 proteins. Human, rhesus, chimpanzee, and African green monkey A3C proteins reduced the relative infectivity of HIV-1Δ*vif* similarly by approximately 60–70%. Conversely, smmA3C-like protein inhibited HIV-1Δ*vif* replication by more than one order of magnitude (Figure 1(a)). Human A3G served as a positive control for major anti-HIV-1 activity. Viral vector-producing cells showed that expression levels of smmA3C-like protein and agmA3C were lower than those of A3Cs from human, rhesus, and cpz (Figure 1(b)). Efficiency of viral incorporation of the smmA3C-like protein was similar to that of hA3G, but much lower compared to hA3C (Suppl. Figure S1(a)).

The smmA3C-like construct was originally described to express A3C of sooty mangabey monkey.[69] However, using alignments of primate A3Z2 and related A3 proteins, we found that the open reading frame consists of exons from both smm*A3C* and smm*A3F* genes. We fused these exons during the PCR amplification step, which occurred because of the high sequence similarity and poor annotation of the smm genome (see discussion section). In the smmA3C-like construct, first (coding for amino acids ^{1}MNPQIR^{6}) and last "exon" (amino acids ^{153}FKYC to EILE^{190}) were derived from smm*A3C* (i.e, coding regions of exon 1 and exon 4 of the smm*A3C* gene) while second (amino acids ^{7}NPMK to FRNQ^{58}) and third "exon" (amino acids ^{59}VDPE to GYED^{152}) in smmA3C-like were of smm*A3F* origin (smmA3F C-terminal domain, CTD, exon 5 and exon 6 of smm*A3F* gene) (Suppl. Figure S1(b)). To compare the deamination activity of smmA3C-like to the WT proteins, we cloned the genuine smmA3C and smmA3F-CTD. Immunoblot analysis of cell lysates confirmed that cellular expression of smmA3C-like and smmA3C (WT) were comparable, but the smmA3F-CTD construct failed to yield detectable levels of protein in transfected cells (Figure 1(c)). In contrast to our expectations, only the smmA3C-like protein and not smmA3C showed enhanced cytidine deaminase activity (Figure 1(d)). Not surprisingly, like hA3C[70] smmA3C-like protein formed intracellular RNAse resistant oligomers or high molecular mass (HMM) complexes and did not self-associate in the cytosol (data not shown).

Because restriction of HIV-1Δ*vif* by smmA3C-like protein was similar to or slightly stronger than restriction by hA3G (Figure 1(a)), we analyzed the DNA-editing capacity of these A3s during infection by "3D-PCR".[70,71] DNA sequences in which cytosines are deaminated by A3 activity contain fewer GC base pairs than non-edited DNA, resulting in a lower melting temperature than the original, non-edited DNA. Therefore, successful PCR amplification at lower denaturation temperatures (T_d) (83.5–87.6 °C) by 3D-PCR indicates the presence of A3-edited sequences. 3D-PCR amplification of viral genomic cDNA with samples of cells infected with HIV-1Δ*vif* viruses encapsidating hA3C, rhA3C, cpzA3C, or agmA3C yielded amplicons at $T_d \geq 86.3$ °C, whereas the activity of smmA3C-like protein allowed to produce amplicons at $T_d < 84.2$ °C. In control reactions using virions produced in the presence of hA3G, PCR amplification of viral DNA was detectable at lower T_d (85.2 °C and weakly at 84.2 °C) (Figure 1(e)). Importantly, using the vector control sample (no A3), PCR amplicons could be amplified only at higher T_d (87.6 °C). To study the effect of smmA3C-like protein in HIV-1Δ*vif*, PCR products generated on smmA3C-like protein-edited samples formed at 84.2 °C were cloned and independent clones were sequenced. The novel smmA3C-like protein caused hypermutation in HIV-1Δ*vif* with a rate of 17.16% and predominantly favored the expected GA dinucleotide context (Suppl. Figure S2(a)). Thus, smmA3C-like protein caused a higher G → A mutation rate in HIV-1Δ*vif* than our previously described enhanced activity mutant A3C.S61P (see Figure 2(a) and (b) for sequence and structure), A3G and A3F.[70] In addition, we applied qualitative *in vitro* cytidine deamination assays using A3 proteins isolated from HIV-1Δ*vif* and SIVagmΔ*vif* viral particles.[72,73] This PCR-based assay depends on the sequence change caused by A3s converting a dC → dU in an 80-nucleotide (nt) ssDNA substrate harboring the A3C-specific TTCA motif. Catalytic deamination of dC → dU by A3C is then followed by a PCR that replaces dU by dT generating an MseI restriction site. The efficiency of MseI digestion was monitored by using a similar 80-nt substrate-containing dU instead of dC in the recognition site. As expected, encapsidation of hA3C and hA3C.S61P into the HIV-1Δ*vif* particles, did not yield a substantial product resulting from ssDNA cytidine deamination,[70] whereas smmA3C-like protein generated high amounts of deamination products (Figure 1(f)). Using smmA3C-like protein, the deamination products were observed even after transfection of 10-fold smaller amounts of expression plasmid during virus production. In contrast, A3C and A3C.S61P proteins isolated from SIVagmΔ*vif* particles produced the expected deamination products, whereas smmA3C-like protein exhibited the strongest catalytic activity, regardless of whether encapsidated in SIVagmΔ*vif* or HIV-1Δ*vif* particles (Figure 1(f)). Taken together, we conclude that smmA3C-like protein inhibits HIV-1 by cytidine deamination causing hypermutation of the viral DNA.


A
HIV-1 Δvif
Relative infectivity (%)
100.0
10.0
1.0
***
vector
hA3C
rhA3C
smmA3C-like
cpzA3C
agmA3C
hA3G
B
kDa
55
40
35
25
α HA (A3)
α p24 (CA)
α GAPDH
C
smmA3F-CTD
D
U
S
P
E
TdºC
83.5 84.2 85.2 86.3 87.6
F
smmA3C-like
hA3C
hA3C.S61P
U mock
1 10
HIV-1Δvif
SIVagmΔvif

Identification of the regulatory domain of smmA3C-like protein that mediates HIV-1 restriction

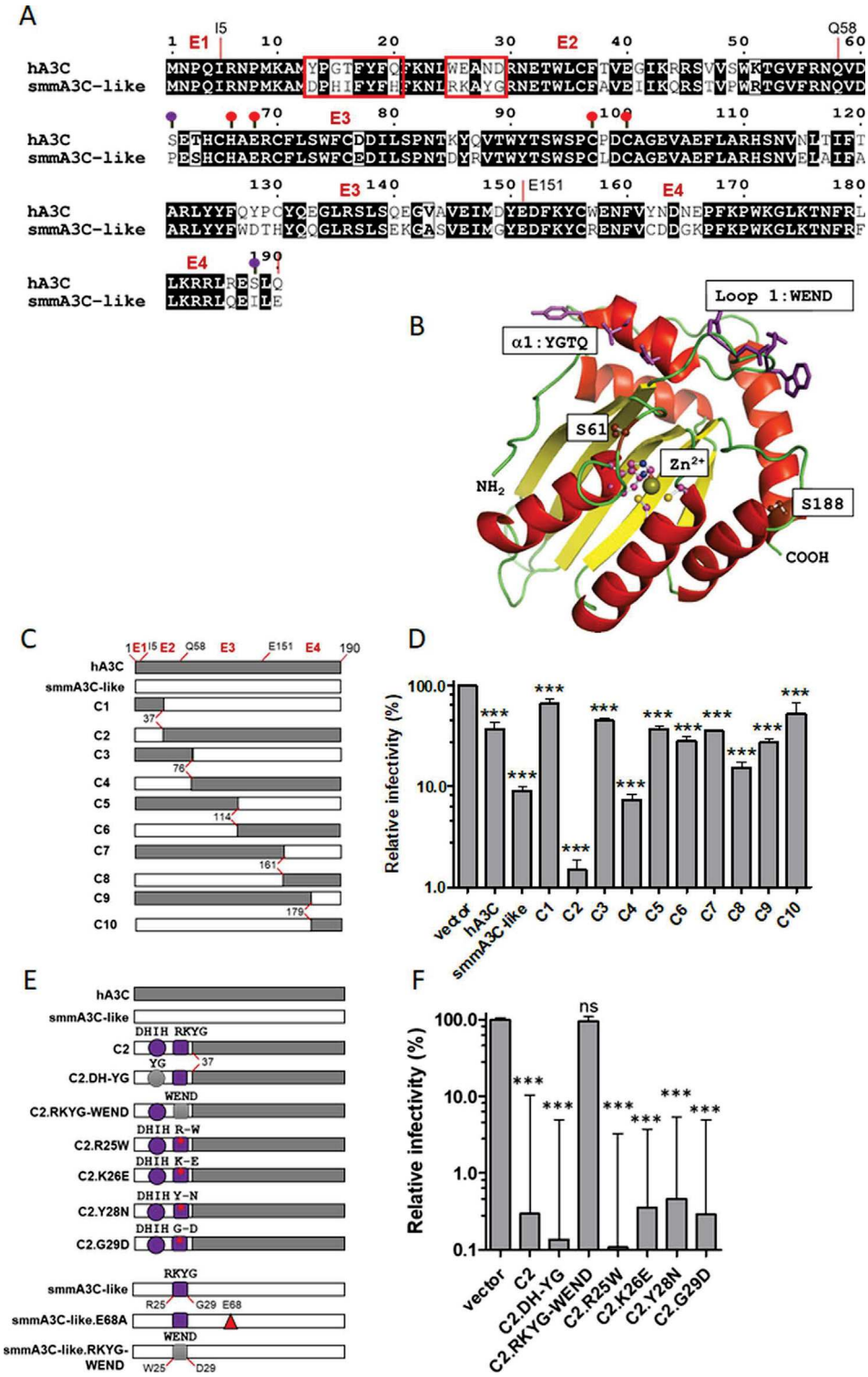
Amino acid sequence identity and similarity between hA3C and smmA3C-like protein reach 77.9% and 90%, respectively (Figure 2(a)). To facilitate the identification of distinct determinants of smmA3C-like protein that confer HIV-1 inhibition, ten different hA3C/smmA3C-like chimeras were constructed (Figure 2(c)).[69] Next, viral particles containing different chimeric proteins were produced and their infectivity was tested. As shown in Figure 2(d), chimeras C2, C4, and C8 strongly reduced the infectivity of HIV-1Δ*vif*. Especially, chimera C2 (hA3C harboring a swap of 36 residues of the smmA3C-like protein at the N-terminal end) inhibited HIV-1Δ*vif* replication by about two orders of magnitude. In comparison, chimeras C6 and C9 reduced viral infectivity by only 72% relative to vector control (Figure 2(d)).

Next, we determined the intracellular expression and virion incorporation efficiency of the chimeras by immunoblot analysis. Chimeras C2, C3, C5, C7, and C9, which contain residues 37 to 76 of hA3C (Figure 2(c)), were more highly expressed than C1, C4, C6, and C10 (Suppl. Figure S2(b)). Specifically, chimera C2 displayed higher protein levels than hA3C while C10 protein was below the detection threshold. Chimeras, C2, C4, C6, C7, and C9 were found to be encapsidated in HIV-1Δ*vif* (Suppl. Figure S2(b), viral lysate). In particular, C3 and C5 were less efficiently packaged into viral particles although they were present at higher intracellular expression levels. Conversely, C6 produced less protein but its viral incorporation was higher than that of C3 or C5. In addition, we analyzed the *in vitro* cytidine deaminase activity of these chimeras as described above (Suppl. Figure S2(c)). Here we used lysates of transfected HEK293T cells to readily evaluate the catalytic activity of the chimeric A3Cs. Only chimers C2 and C4 showed the level of deamination similar to those produced by smmA3C-like protein (Suppl. Figure S2(c)). Taken together, chimeras C2 and C4 have the strongest HIV-1Δ*vif*-restricting effect among all tested chimeras and display corresponding *in vitro* deamination activity. Due to its superior antiviral activity, we mainly focused on chimera C2 in our following experiments.

Synergistic effects of residues in the RKYG motif of chimera C2 and smmA3C-like protein control their potent antiviral activity

To identify the specific residues in chimera C2 that are essential for its anti-HIV-1 activity, we targeted two N-terminal motifs of C2, namely ^{13}DPHIFYFH^{20} (shortly "DHIH") and ^{25}RKAYG^{29} (named "RKYG") as presented in the sequence alignments of Figure 2(a), and generated variants of C2 by swapping one, two, or four amino acids with the analogous residues of hA3C as presented in Figure 2(e). First, we cloned the C2 variants C2.DH-YG (YGTQ motif of helix α1) and C2.RKYG-WEND (WEND motif of loop 1, Figure 2 (a) and (b)) and tested their anti-HIV-1 and deamination activity. This pilot experiment revealed that loop 1 motif RKYG but not α1 helix motif DHIH in C2 is essential for its activity (Figure 2(f) and Suppl. Figure S3(a)). Hence, we constructed the mutants C2.R25W, C2.K26E, C2.Y28N, and C2.G29D (Figure 2(e)) and tested them for catalytic and antiviral activity. The results of the deamination assay further demonstrated that the DH motif in C2 is not relevant for its potent catalytic activity, as the C2.DH-YG acted

Figure 1. A3C-like protein from sooty mangabey inhibits HIV1Δ*vif* by more than 10-fold. (a) HIV-1Δ*vif* particles were produced with A3C from human, rhesus macaque, chimpanzees (cpz), African green monkey (agm), and A3C-like protein from sooty mangabey monkey (smm), hA3G, or vector only. Infectivity of equal amounts of viruses (RT-activity normalized), relative to the virus lacking any A3, was determined by quantification of luciferase activity in HEK293T cells. Presented values represent means ± standard deviations (error bars) for three independent experiments. Asterisks indicate statistically significant differences relative to the effect of the empty vector on infectivity: ***, $p < 0.0001$. (b) Immunoblot analysis of HA-tagged A3 and HIV-1 capsid expression in cell lysates using anti-HA and anti p24 antibodies, respectively. GAPDH served as a loading control. "α" represents anti. (c) Expression and (d) deamination activity of smmA3C and smmA3F-CTD: Immunoblot analysis of HA-tagged smmA3C-like protein, and (WT) smmA3C, and (WT) smmA3F-CTD expression in cell lysates using anti-HA antibody. Tubulin served as a loading control. *In vitro* deamination activity of smmA3C-like protein, smmA3C, and smmA3F-CTD using lysates of cells that were previously transfected with the respective expression plasmids. Samples were treated with RNAse A; oligonucleotide-containing uracil (U) instead of cytosine served as a marker to denote the migration of deaminated product after restriction enzyme cleavage. S-substrate, P-product. (e) 3D-PCR: HIV-1Δ*vif* produced together with A3C orthologues, hA3G or vector controls were used to transduce HEK293T cells. Total DNA was extracted and a 714-bp fragment of reporter viral DNA was selectively amplified using 3D-PCR. T_d = denaturation temperature. Extensive viral DNA editing profile of smmA3C-like protein and its relative positions of G → A transition mutations are presented in Suppl. Figure S2(a). (f) *In vitro* deamination activity of A3Cs encapsidated in HIV-1Δ*vif*, and SIVagmΔ*vif* particles. Virions were concentrated and lysed in mild lysis buffer and equal amounts of lysate were used for the assay. Numbers 1 and 10 indicate 60 ng and 600 ng of A3 expression vector used for transfection, respectively.


A
hA3C
smmA3C-like
E1
I5
E2
E3
E4
Q58
E151
B
Loop 1:WEND
α1:YGTQ
S61
Zn2+
NH2
S188
COOH
C
hA3C
smmA3C-like
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
D
Relative infectivity (%)
100.0
10.0
1.0
vector
hA3C
smmA3C-like
E
hA3C
smmA3C-like
C2
C2.DH-YG
C2.RKYG-WEND
C2.R25W
C2.K26E
C2.Y28N
C2.G29D
smmA3C-like.E68A
smmA3C-like.RKYG-WEND
F
Relative infectivity (%)
100.0
10.0
1.0
0.1
ns
vector
C2
C2.DH-YG
C2.RKYG-WEND
C2.R25W
C2.K26E
C2.Y28N
C2.G29D

similar to C2 (Suppl. Figure S3(a)), but mutation of the RKYG motif in the RKYG-WEND variant resulted in a loss of deamination activity (Suppl. Figure S3(a)). Interestingly, none of the single amino acid changes in RKYG (R25W, K26E, Y28N, and G29D) resulted in the loss-of-function of C2, albeit the catalytic activities of R25W and K26E were partially reduced (Suppl. Figure S3 (a)). Consistent with the data obtained from the *in vitro* assay, the chimeric C2.RKYG-WEND variant failed to restrict the infectivity of HIV-1Δ*vif* (Figure 2(f)). Immunoblot analysis of cell and viral lysates confirmed that cellular expression and viral encapsidation of these variants were comparable (Suppl. Figure S3(b)). Finally, to test the *in vivo* DNA-editing capacity, we performed 3D-PCR analysis using C2, C2.DH-YG, and C2. RKYG-WEND variants. As presented in the 3D-PCR experiment of Suppl. Figure S3(c), only HIV-1Δ*vif* particles produced in the presence of A3C chimera C2 and its mutant C2.DH-YG generated amplicons that were detected at low-denaturation temperature, and C2.RKYG-WEND behaved similar to the vector control. Likewise, replacing RKYG with WEND in the smmA3C-like protein (Figure 2(e)) inhibited its antiviral activity (Figure 3 (a) and (b)), DNA-editing capacity of HIV-1 genomes (Figure 3(c)), and catalytic activity *in vitro* (Figure 3(d)) as did the active site mutant E68A.

The WE-RK mutation in loop 1 of hA3C determines its strong deaminase-dependent antiviral function

Mutational changes of the RKYG motif to WEND residues in loop 1 of C2 and smmA3C-like protein resulted in complete loss of enzymatic functions and anti-HIV-1 activities (Figures 2(f), 3(a), (c), (d), and Suppl. Figs. S3(a) and (c)). To identify the residues in hA3C that are critically required for the deaminase-dependent antiviral activity against HIV-1Δ*vif*, we mutated the loop 1 of hA3C with ^{25}WE^{26} to ^{25}RK^{26} and ^{28}ND^{29} to ^{28}YG^{29} residues and compared their antiviral capacity (see A3C alignment and ribbon diagram Figure 2(a) and (b)). As controls, we included additional mutants such as a catalytically inactive non-Zn^{2+}-coordinating C97 mutant, A3C.C97S,[57] and the variants A3C.S61P[70] and A3C.S188I[74] exhibiting enhanced deaminase activity. Compared to WT hA3C, WE-RK greatly enhanced inhibition of HIV-1Δ*vif* and the ND-YG variant behaved like WT A3C, while S61P and S188I demonstrated only marginally increased HIV-1Δ*vif* restriction (Figure 4 (a)). Importantly, active site mutant A3C.C97S did not inhibit HIV-1Δ*vif* (Figure 4(a)). Enhancement of the antiviral activity of hA3C.WE-RK compared to WT hA3C appear to result neither from higher protein expression in the virus producer cells nor from differences in encapsidation, as demonstrated in a titration experiment that directly compared these features for both proteins (Suppl. Figure S4 (a)).

Next, we asked if the antiviral activity of A3C.WE-RK is deamination-dependent. To achieve this, we introduced the C97S mutation in A3C.WE-RK. Additionally, we compared the ancillary effect of mutants such as S61P[70] and S188I[74] by introducing these mutations in the WE-RK variant of A3C. As expected, the inhibitory activities of A3C.WE-RK, A3C.WE-RK.S61P, and A3C.WE-RK.S61P. S188I against HIV-1Δ*vif* were abolished by the active site ablating mutation C97S, indicating the importance of the enzymatic activity of A3C (Figure 4(b)). In comparison, introducing either the

Figure 2. Design and activity of hA3C/smmA3C-like protein chimeras. (a) Sequence alignment of hA3C and smmA3C-like protein, motif 1 (YGTQ) and motif 2 (WEND) are marked with red boxes; Red lollipops indicate active site amino acids H66, E68, C97 and C100, while S61 and S188 are colored in purple. (b) Ribbon model of the crystal structure of A3C (PDB 3VOW) depicting the spatial arrangements of helix α1 (YGTQ motif) and loop 1 (WEND motif). Residues of both motifs are presented in purple. Key residues S61, S188, and zinc-coordinating active site residues are denoted as ball and sticks. Sphere represents Zn^{2+} ion. (c) Structures of the chimeras generated between A3C and smmA3C-like protein. Grey and white boxes indicate fractions of A3C and the smmA3C-like protein, respectively. Regions of hA3C protein derived from exons E1 (amino acids 1–5), E2 (6–58), E3 (59–151), and E4 (152–190) and residues at the borders are marked on top of the hA3C box. Each chimera ("C") encompasses 190 amino acids. Amino acid position (number) at the breakpoints of each chimera is indicated. (d) HIV-1Δ*vif* particles were produced with A3C from human, smm (A3C-like), and h/smm chimeras or vector only. Infectivity of equal amounts of viruses (RT-activity normalized), relative to the virus lacking any A3, was determined by quantification of luciferase activity in HEK293T cells. (e) Illustration of chimera 2 (C2) and variants of C2 or smmA3C-like protein having amino acid exchanges in the DHIH (circle) or RKYG (square) motif. The red triangle denotes catalytic residue E68A mutation. Amino acid position (number) at the breakpoint of chimera C2 is indicated. (f) HIV-1Δ*vif* particles were produced with C2 and its variants or vector only. Infectivity of equal amounts of viruses (RT-activity normalized), relative to the virus lacking any A3, was determined by quantification of luciferase activity in HEK293T cells. Values are means ± standard deviations (error bars) for three independent experiments. Presented values represent means ± standard deviations (error bars) for three independent experiments. Asterisks indicate statistically significant differences relative to the effect of the empty vector on infectivity: ***, $p < 0.0001$; ns, not significant.

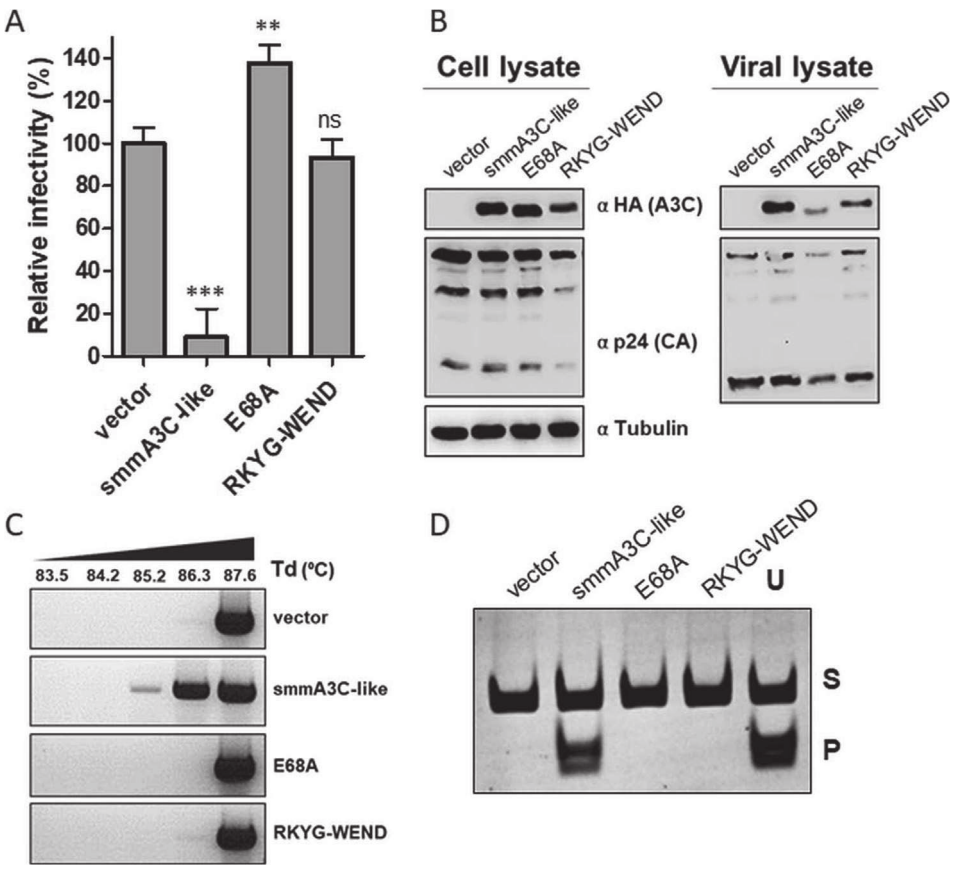
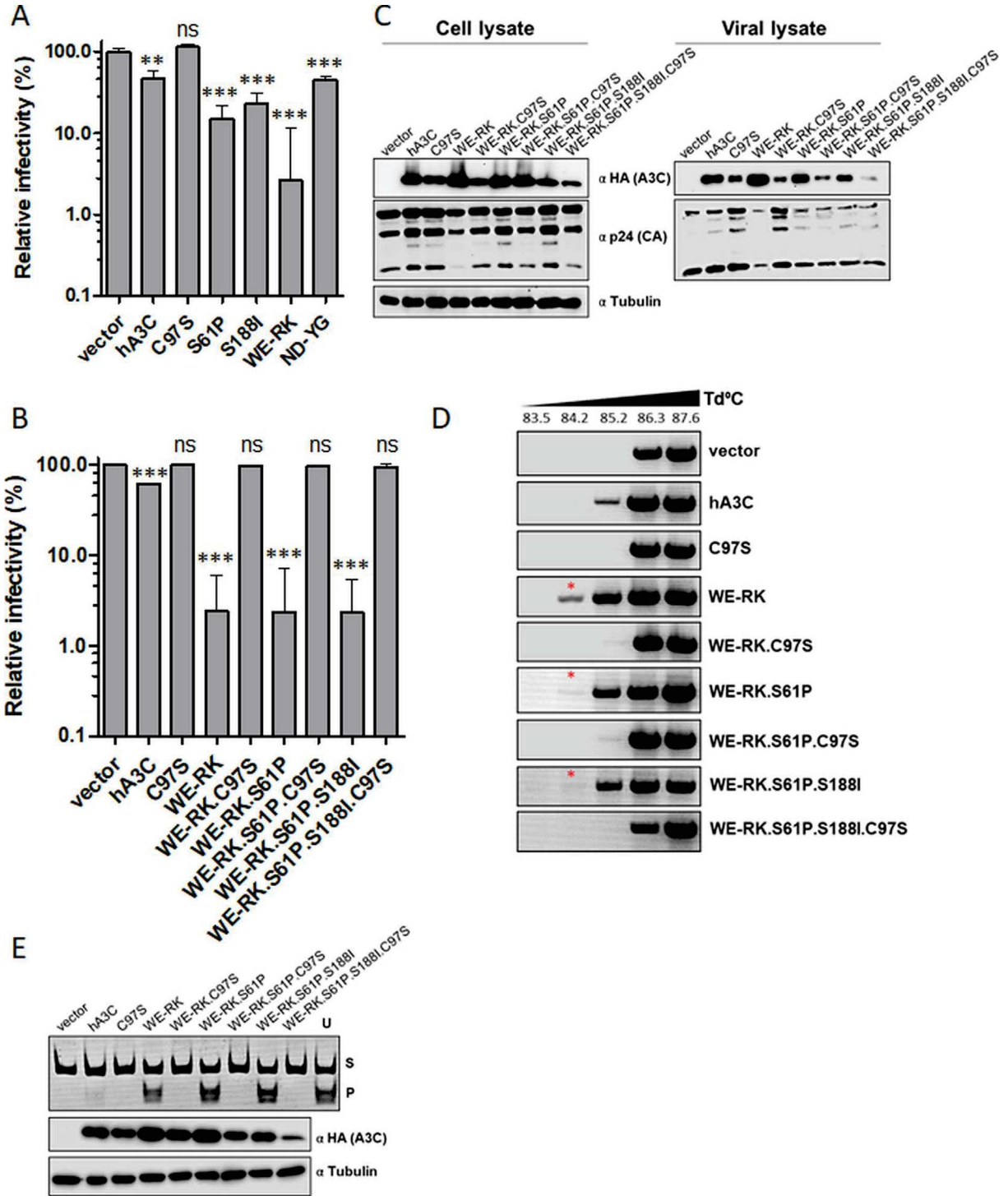


Figure 3. RKYG-WEND exchange in smmA3C-like protein abrogates its antiviral activity. (a) HIV-1Δ*vif* particles were produced with smmA3C-like protein, its mutants E68A (catalytically inactive), RKYG-WEND or vector only. Infectivity of equal amounts of viruses (RT-activity normalized) relative to the virus lacking any A3 was determined by quantification of luciferase activity in HEK293T cells. (b) Immunoblot analyses were performed to quantify HA-tagged A3 proteins and viral p24 proteins in cellular and viral lysates using anti-HA and anti-p24 antibodies, respectively. Tubulin served as a loading control. "α" represents anti. (c) Quantification of hypermutation in viral DNA by 3D-PCR. HIV-1Δ*vif* particles produced in the presence of overexpressed smmA3C-like protein, its variants or vector control were used to transduce HEK293T cells. Total DNA was extracted and a 714-bp fragment of reporter viral DNA was selectively amplified using 3D-PCR. T_d = denaturation temperature. (d) *In vitro* deamination activity of smmA3C-like protein and its variants using lysates of cells that were previously transfected with the respective expression plasmids. Samples were treated with RNAse A; oligonucleotide-containing uracil (U) instead of cytosine served as a marker to denote the migration of deaminated product after restriction enzyme cleavage. S-substrate, P-product.

single mutation S61P or the double mutation S61P. S188I did not considerably change the activity of A3C.WE-RK (Figure 4(b)). Immunoblot analysis of cell and viral lysates demonstrated that hA3C and all mutants (except A3C.WE-RK.S61P.S188I.C97S mutant) were expressed at comparable levels (Figure 4(c)). However, viral incorporation of A3C. C97S, A3C.WE-RK.C97S, A3C.WE-RK.S61P. C97S, and WE-RK.S61P.S188I.C97S was slightly decreased relative to that of WT and mutant proteins that do not contain the C97S mutation (Figure 4(c)). Moreover, we confirmed the effects of all mutants on HIV-1Δ*vif* propagation by 3D-PCR (Figure 4(d)) and deamination assays *in vitro* (Figure 4(e)). In both assays, we found that the C97S mutation destroyed the function of all A3C variants. Thus, we conclude that the loop 1-mediated enhanced activity of hA3C.WE-RK is dependent on catalytic deamination.

To address if the cellular localization of A3C is affected by the WE-RK mutations, we used confocal microscopy. HeLa cells were transfected with the HA-tagged hA3C or hA3C.WE-RK and the proteins were visualized by applying an anti-HA antibody. Both proteins, hA3C and hA3C.WE-RK were localized in cytoplasm and nucleus (Figure 5(a) and (b)). This distribution was found in 65.5% and 75% of cells expressing hA3C or that of hA3C.WE-RK, respectively. Only 20% or 10% of the cells expressing hA3C or hA3C.WE-RK, respectively, displayed these proteins solely in the nucleus (Figure 5(c)). Together, we infer that hA3C and hA3C.WE-RK had a similar distribution in HeLa cells.


A
Relative infectivity (%)
100.0
10.0
1.0
0.1
ns
**
***
vector
hA3C
C97S
S61P
S188I
WE-RK
ND-YG
C
Cell lysate
Viral lysate
vector
hA3C
C97S
WE-RK
WE-RK.C97S
WE-RK.S61P
WE-RK.S61P.C97S
WE-RK.S61P.S188I
WE-RK.S61P.S188I.C97S
α HA (A3C)
α p24 (CA)
α Tubulin
B
Relative infectivity (%)
100.0
10.0
1.0
0.1
ns
***
vector
hA3C
C97S
WE-RK
WE-RK.C97S
WE-RK.S61P
WE-RK.S61P.C97S
WE-RK.S61P.S188I
WE-RK.S61P.S188I.C97S
D
TdºC
83.5 84.2 85.2 86.3 87.6
vector
hA3C
C97S
WE-RK
WE-RK.C97S
WE-RK.S61P
WE-RK.S61P.C97S
WE-RK.S61P.S188I
WE-RK.S61P.S188I.C97S
E
vector
hA3C
C97S
WE-RK
WE-RK.C97S
WE-RK.S61P
WE-RK.S61P.C97S
WE-RK.S61P.S188I
WE-RK.S61P.S188I.C97S
U
S
P
α HA (A3C)
α Tubulin

The RK-WE mutation in loop 1 moderately reduces the antiviral activity of hA3F

hA3C and hA3F-CTD display 77% sequence similarity, reflecting a common evolutionary origin.[6] Interestingly, the antiviral activity of hA3F is mediated by its CTD.[75,76] Various loops within the A3F-CTD were recently investigated for their role in substrate binding and enzyme function[77] but it was not possible to unravel the antiviral activity of a protein consisting only of the A3F-CTD, mainly due to earlier reported difficulties in expressing this domain alone in human cells.[70,78] The residues ^{25}RK^{26} in loop 1 of smmA3C-like protein are derived from exon 5 of the smm*A3F* gene, located in the CTD of A3F (Suppl. Figure S1(b)) and are conserved in primate A3F proteins (see section evolution, below). To test the impact of RK residues in CTD loop 1 of the hA3F, we compared the antiviral activity of hA3F with A3F.RK-WE against HIV-1Δ*vif*. hA3F and hA3F.RK-WE yielded similar amounts of protein and were equally efficiently encapsidated in HIV-1 particles (Figure 6(a)). However, the HIV-1Δ*vif* inhibiting effect of A3F.RK-WE was about 2-fold lower than WT A3F (Figure 6(b)). Consequently, A3F.RK-WE showed decreased mutation efficiency compared with WT A3F (Figure 6(c) and (d)), which is consistent with data presented in a recent report.[77] Thus, we conclude that loop 1 with its residues RK in CTD of A3F is important for the enzymatic function of hA3F.

Inhibition of human LINE-1 retrotransposition by A3C variants

Since A3C and A3F restrict endogenous human LINE-1 (L1) retrotransposition activity by 40–75% and 66–85%, respectively,[51,61,79,80] we set out to elucidate how the WE and the RK residues in loop 1 of both hA3C and hA3F affect the L1 inhibiting activity. To this end, we quantified the L1-inhibiting effect of human WT A3A, A3C, and A3F proteins and their mutants hA3C.WE-RK, hA3C.WE-RK.S61P, and hA3F.RK-WE applying a dual-luciferase retrotransposition reporter assay.[81] In this cell culture-based assay, the firefly luciferase gene is used as the reporter for L1 retrotransposition and the Renilla luciferase gene is encoded on the same plasmid for transfection normalization (Figure 7(a)). Consistent with previous reports[51,61], overexpression of hA3A, hA3C, and hA3F inhibited L1 reporter retrotransposition by approximately 94%, 68%, and 56%, respectively (Figure 7(b)). The mutant hA3C.WE-RK restricted L1 more strongly (from 56% to ~96%), but the introduction of the additional S61P mutation in hA3C.WE-RK.S61P did not further increase the ability of the enzyme to restrict L1 mobilization (Figure 7(b)). Notably, hA3F and the mutant hA3F.RK-WE exhibited a comparable level of L1 restriction, indicating that regions other than loop 1 of A3F-CTD and, probably, the NTD (N-terminal domain) of hA3F are involved in L1 restriction (Figure 7(b)). Immunoblot analysis of cell lysates of co-transfected HeLa-HA cells demonstrated comparable expression of the L1 reporter and HA-tagged A3 and A3 mutant proteins (Suppl. Figure S4(b)). Furthermore, compared to the inhibition of L1 retrotransposition by hA3C and chimpanzee A3C (~60%), hA3C.S61P inhibited L1 reporter retrotransposition by 75% (Suppl. Figure S4(c) and (d)). These findings indicate that the WE-RK mutation in hA3C enhances its L1-inhibiting activity. Based on the observed antiviral activity and the L1-restricting effect of hA3C.WE-RK on L1, we hypothesize that the introduction of these positively charged residues in hA3C significantly fosters its interaction with nucleic acids, which was recently reported to mediate its L1 inhibiting activity.[61]

Figure 4. A3C gains deaminase-dependent anti-HIV-1 activity by a WE-RK change in loop 1. (a) HIV-1Δ*vif* particles were produced with hA3C, its mutants (C97S, S61P, S188I, WE-RK, ND-YG), or vector only. Infectivity of equal amounts of viruses (RT-activity normalized), relative to the virus lacking any A3C, was determined by quantification of luciferase activity in HEK293T cells. (b) HIV-1Δ*vif* particles were produced with hA3C, its variants such as C97S, WE-RK, WE-RK.C97S, WE-RK.S61P, WE-RK.S61P.C97S, WE-RK.S61P.S188I, WE-RK.S61P.S188I.C97S or vector only. Infectivity of equal amounts of viruses (RT-activity normalized), relative to the virus lacking any A3C, was determined by quantification of luciferase activity in HEK293T cells. (c) Quantification of HA-tagged WT and mutant A3C proteins in both cellular and viral lysates by immunoblot analysis. A3s and HIV-1 capsids were stained with anti-HA and anti-p24 antibodies, respectively. Tubulin served as a loading control. "α" represents anti. (d) 3D-PCR: HIV-1Δ*vif* produced together with hA3C, its variants (as in (b)), or vector controls were used to transduce HEK293T cells. Total DNA was extracted and a 714-bp fragment of reporter viral DNA was selectively amplified using 3D-PCR. T_d = denaturation. (e) *In vitro* deamination assays to examine the catalytic activity of A3C and its variants using lysates of cells that were previously transfected with respective expression plasmids (as in (b)). RNAse A-treatment was included; oligonucleotide containing uracil (U) instead of cytosine served as a marker to denote the migration of the deaminated product after restriction enzyme cleavage. S-substrate, P-product. The two lower panels represent immunoblot analyses of expression levels of HA-tagged A3C and mutant proteins (α HA (A3C)) and tubulin (α tubulin), which was used as a loading control.

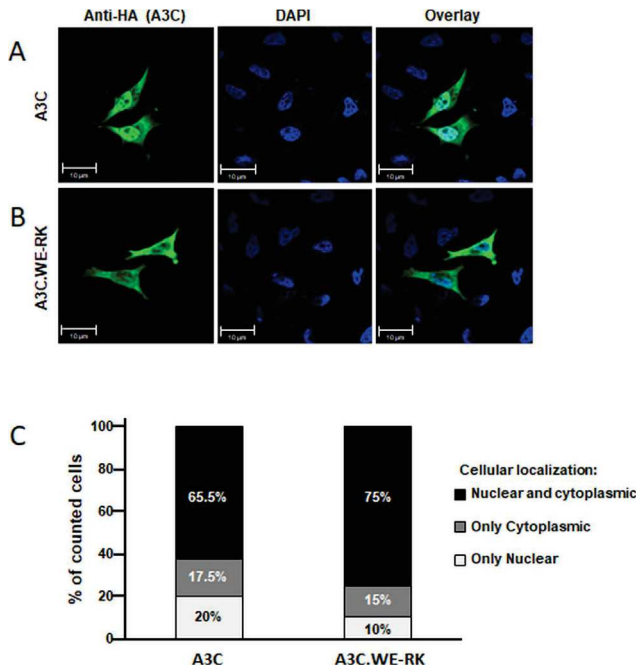


Figure 5. Subcellular localization of human A3C in transfected HeLa cells. Immunofluorescence confocal laser scanning microscopy images of HeLa cells transfected with HA-tagged A3C or A3C.WE-RK. Representative pictures are shown which illustrate nuclear and cytoplasmic localization of the A3Cs ((a) and (b)) x-y optical sections. To detect A3Cs (green) immunofluorescence, cells were stained with an anti-HA antibody. Nuclei (blue) were visualized by DAPI staining. (c) 40 randomly chosen transfected cells with A3C or A3C.WE-RK were categorized and cellular localization of A3Cs were quantified.

The positively charged residues R25 and K26 in A3C form salt-bridges with the backbone of the ssDNA

To understand how the positively charged residues in loop 1 of A3C.WE-RK mediate the enhanced cytidine deamination activity, a structural model of hA3C variant hA3C-RKYG binding to ssDNA, based on the ssDNA-bound crystal structure of A3A was generated that shows a cytidine residue in the active center of hA3C. RKYG (Suppl. Figure S5(a)). However, the ssDNA fragment (which was co-crystallized with hA3A) in this conformation is too short to interact with amino acids 25, 26, 28, and 29, which differ between hA3C WT and the hA3C.RKYG variant. Hence, this static binding mode model cannot explain why hA3C.RKYG has a higher cytidine deaminase activity than hA3C WT. To probe the impact of structural dynamics on residue-ssDNA interactions in order to explain the differences in A3C.WE-RK properties, this model was later subjected to molecular dynamics (MD) simulations.

To assess the binding to a longer ssDNA fragment, we generated a second complex model of ssDNA bound to the NTD of rhesus macaque A3G (rhA3G),[82] similar to the ssDNA-bound A3F-CTD model built previously;[83] and aligned the crystal structure of hA3C WT and the model of hA3C. RKYG to this complex (Figure 8(a)–(c). Note that the A3G structure was used only for placing the DNA but not for modeling the protein part). This new model revealed that the positively charged residues R25 and K26 in hA3C.RKYG form salt-bridges with the backbone of the ssDNA (Figure 8 (c)) in contrast to hA3C WT (Figure 8(b)). Thus, these two residues can form stronger interactions with ssDNA in hA3C.RKYG than their counterparts in hA3C, which may explain the enhanced cytidine deaminase activity of hA3C.WE-RK compared to hA3C (Figure 4(e)). However, as the binding of ssDNA to NTDs, such as in the structure of rhA3G, differs from that in CTDs, we did not subject the former model to MD simulations.

We next performed five replicas of MD simulations of 2 μs length each for hA3C, hA3C.RKYG, and hA3C.S61P.S188I to assess the structural impact of the substitutions on the protein. For this purpose, we used a hA3C crystal structure as starting structures and variants thereof generated by substituting respective residues. In all MD simulations, the cytidine remains bound to the Zn^{2+} ion in the active site. The root mean square fluctuations (RMSF), which describe atomic mobilities during the MD simulations, show distinct differences between the variants in the putative DNA-binding regions of the proteins: the RMSF of hA3C.RKYG and hA3C.S61P.S188I are up to 2 Å larger compared to hA3C WT in the regions carrying the substitutions (residues 21–32 for

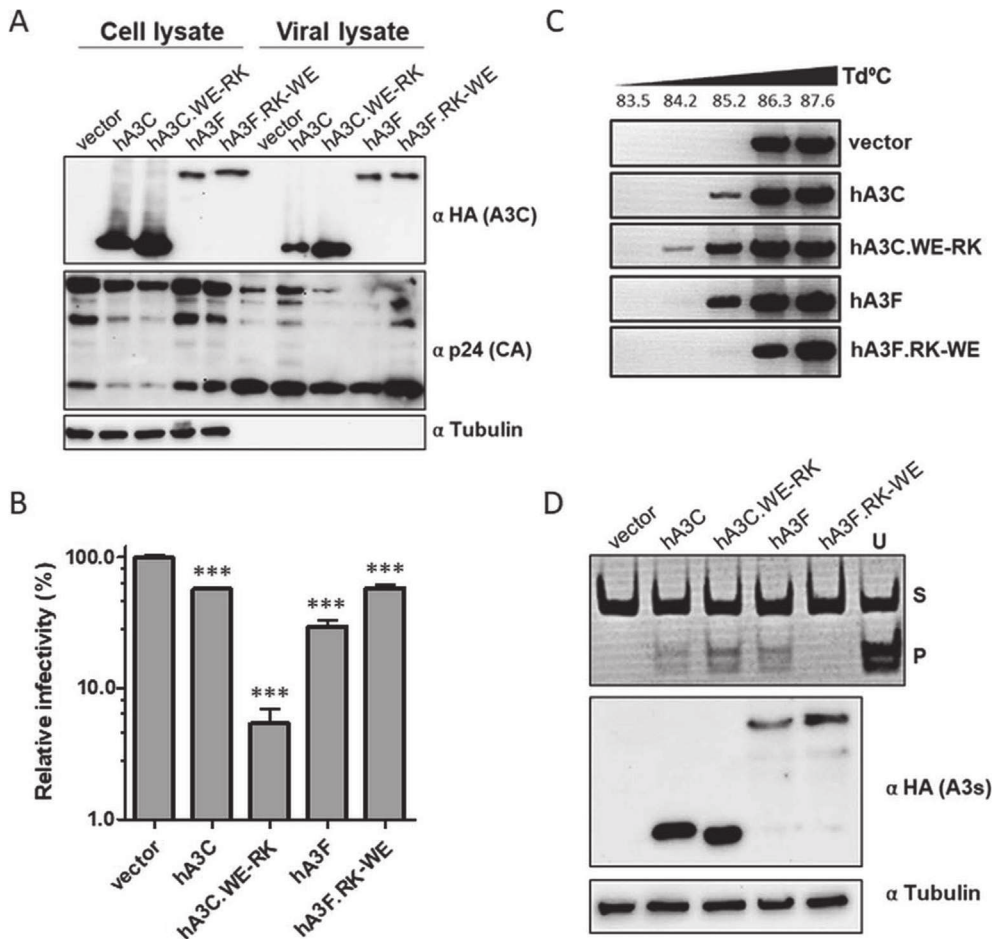


Figure 6. Mutations in loop 1 of A3F-CTD moderately affect the antiviral activity of A3F. (a) Immunoblot analyses were performed to quantify the amounts of HA-tagged WT hA3C and hA3F proteins and their loop 1 mutants in cell lysates and viral particles. HA-tagged A3s and HIV-1 capsid proteins were stained with anti-HA and anti-p24 antibodies, respectively. Tubulin served as a loading control. "α" represents anti. (b) Infectivity of equal amounts of HIV-1Δ*vif* viruses (RT-activity normalized) encapsidating hA3C, hA3F, or their loop 1 mutants relative to the virus lacking any A3 protein was determined by quantification of luciferase activity in transduced HEK293T cells. (c) 3D-PCR: HIV-1Δ*vif* produced together with hA3C, hA3F, and their loop 1 mutants or vector control were used to transduce HEK293T cells. Total DNA was extracted and a 714-bp fragment of reporter viral DNA was selectively amplified using 3D-PCR. T_d = denaturation temperature. (d) *In vitro* deamination assay to examine the catalytic activity of hA3C, hA3F, and their loop variants was performed using lysates of cells that were transfected with the respective A3 expression plasmids. RNAse A-treatment was included; oligonucleotide containing uracil (U) instead of cytosine served as a marker to denote the migration of the deaminated products after restriction enzyme cleavage. S-substrate, P-product. The two lower panels represent immunoblot analyses of expression levels of HA-tagged A3C, A3F and mutant proteins (α HA (A3s)) and tubulin (α tubulin), which was used as a loading control.

hA3C.RKYG and residues 55–67 for hA3C.S61P.S188I) (Suppl. Figure S5(b)). This effect is specifically related to the respective substitutions, as no change in RMSF occurs for a variant in any region where it is identical to A3C WT. The increased movement of ssDNA-binding residues might improve the sliding of hA3C.RKYG and hA3C.S61P.S188I along the ssDNA, owing to more transient interactions with the ssDNA backbone. Conversely, the RMSF of loop 7 is up to 1 Å lower in both the hA3C.RKYG and hA3C.S61P.S188I variants compared to the hA3C WT (Suppl. Figure S5(b)).

These results encouraged us to investigate possible interaction patterns between DNA and each of the three A3C variants that could be a result of the shift in loop 1 dynamics. For this purpose, we used the initial DNA-bound model of hA3C.RKYG with cytidine in the active center, modeled from the experimental A3A structure as described above, to generate DNA-bound complexes for hA3C WT and hA3C.S61P.S188I. While our MD simulations showed similar changes in the conformational dynamics of the loops as before (Suppl. Figure S5(b)), we detected an interesting change in interactions

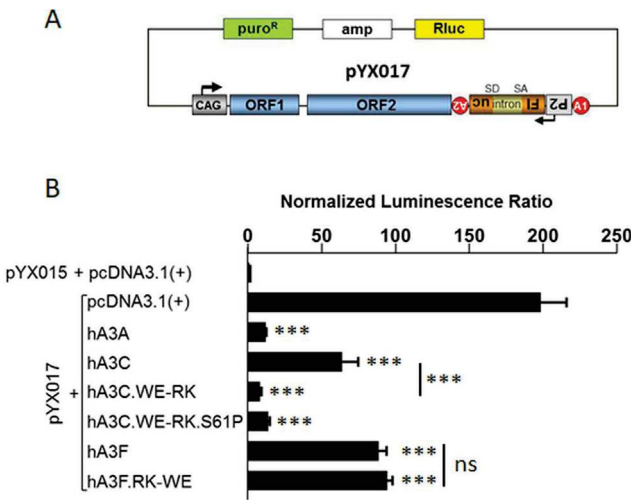


Figure 7. Expression of the hA3C.WE-RK variant enhances A3C-mediated L1 restriction significantly. Dual-luciferase reporter assay to evaluate the effect of WT and mutant A3 proteins on L1 retrotransposition activity. (a) Schematic of the L1 retrotransposition reporter construct pYX017.[81] The $L1_{Rp}$ reporter element is under transcriptional control of the CAG promoter and a polyadenylation signal (A1) at its 3′end. The firefly luciferase (Fluc) cassette has its own promoter (P2) and polyadenylation signal (A2), is expressed from the antisense strand relative to the CAG promoter, and interrupted by an intron (with splice donor [SD] and splice acceptor [SA]) in the transcriptional orientation of the L1 reporter element. (b) Effect of WT and mutant A3 proteins on L1 retrotransposition activity indicated by normalized luminescence ratio (NLR). NLR indicating retrotransposition activity observed after cotransfection of pYX015 and empty pcDNA3.1 (+) expression plasmid was set as 1. Error bars indicate standard deviation ($N = 4$).

between loop 1 residue R30 and the DNA. R30, which is present in all three variants and points away from the DNA in the A3C crystal structure, interacts more frequently with the DNA in both hA3C.S61P.S188I (16.4 ± 2.6% of the simulation time applying stringent criteria for H-bond formation (mean ± SEM for 10 trajectories)) and hA3C.RKYG (44.7 ± 2.7%) than in hA3C WT (0.1 ± 0.0%). In hA3C.RKYG, K26 similarly forms H-bonds with the DNA over 10.3 ± 2.8% of the MD trajectories, but, expectedly, E26 in hA3C WT and hA3C.S61P.S188I forms almost no H-bonds.

In addition, to rule out the possibility that the loop 7 residues might be influencing the loop 1 residues from binding DNA, we have analyzed the interaction between them. The average distance between the two loops in the absence of DNA is very similar for hA3C (12.1 ± 1.75 Å; SD, n = 5000), hA3C.RKYG (12.7 ± 1.78 Å; SD, n = 5000), and hA3C.S61P.S188I (12.12 ± 1.78 Å; SD, n = 5000). Given the average distance of 12 Å it is not surprising that with the exception of N23 and A121, which are the only residues in spatial proximity and thus commonly interact, residues in loop 1 form H-bonds to those in loop 7 in less than 1% of the simulation time for all variants. The average distance between any atom in residue 25 to residues in loop 7 is larger than 4.4 Å, suggesting that sustained interactions are unlikely.

Next, we used more lenient distance criteria suitable to evaluate the formation of interactions and evaluated, whether only the N terminus (W25 in hA3C and hA3C.S61P.S188I and R25 in hA3C.RKYG) or only the C terminus (R30 in all three variants) of loop 1, or both residues at the same time, interact with the DNA. In hA3C, only W25 interacts with the DNA in ~20% of the conformations (Suppl. Figure S6(a)). In hA3C.S61P.S188I, interactions between W25 or R30 occur in ~20% of the conformations, thus showing an increase of a factor of 5 for R30 (Suppl. Figure S6(b)). In hA3C.RKYG, both R25 and R30 simultaneously interact with DNA in ~29% of all investigated conformations besides the interactions of R30 with DNA alone in ~42% of the conformations (Suppl. Figure S6(c)). Hence, these results suggest that W25 and R30 act additively in hA3C.S61P.S188I, whereas they act cooperatively in hA3C.RKYG. This correlates with the differences in activities, with hA3C.RKYG showing the highest activity against HIV-1Δ*vif*.

WE-RK mutation in the loop 1 of A3C enhances the interaction with ssDNA

To validate our structural modeling analysis and to address if the interaction of hA3C and hA3C.WE-RK with the substrate ssDNA was differentially affected, we performed electrophoretic mobility shift assays (EMSA) using hA3C-GST (A3C fused to glutathione S-transferase, GST) and hA3C.WE-RK-GST purified from HEK293T cells (Figure 9(a)). We first

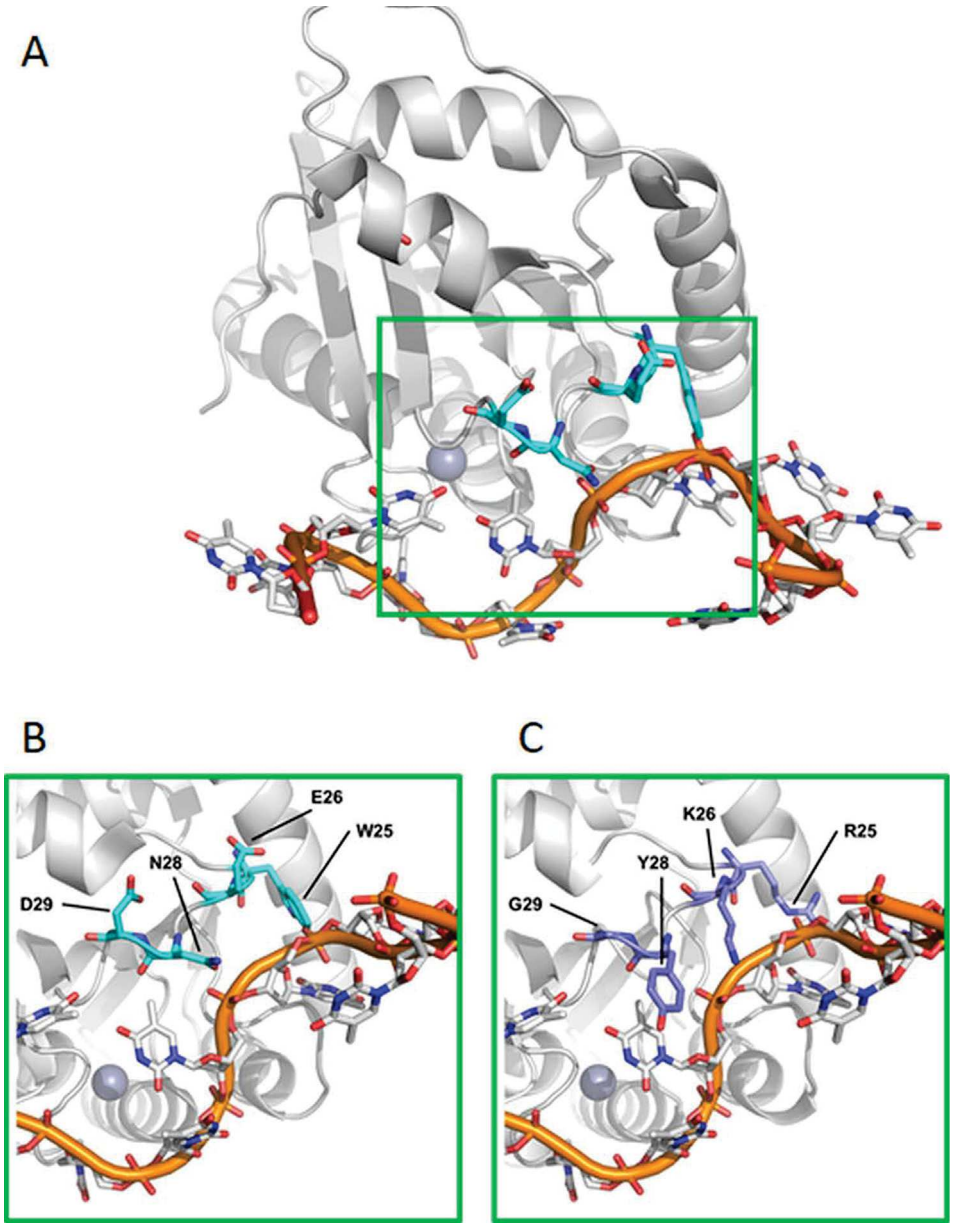


Figure 8. ssDNA-protein interaction model of hA3C and hA3C.RKYG. (a) Binding mode model of ssDNA (orange) to hA3C WT based on hA3F-CTD and rhA3G-NTD. Magnifications of the active center (green box) are shown at the bottom for hA3C WT (b) and hA3C.RKYG (c). The side chains of residues in the active center that differ between hA3C WT and the hA3C.RKYG variant are shown in cyan and dark blue, respectively. The Zn^{2+} ion in the active center is shown as a sphere. Ongoing from hA3C WT to the hA3C.RKYG variant, the interface changes from being negatively to being positively charged. The flexible arginine and lysine side chains in the hA3C.RKYG variant can interact with the negatively charged backbone of ssDNA (panel C), stabilizing this interaction.

confirmed that the purified GST fusion proteins are catalytically active (Figure 9(b)). As expected hA3C.WE-RK-GST displayed a stronger enzymatic activity than the WT equivalent and no activity with GST was detected (Figure 9(b)). For EMSA, as a probe, we used a biotin-labeled ssDNA oligonucleotide that harbors a TTCA motif in its central region.[70,84] Because hA3C-GST is known to form a stable DNA-protein complex when the protein concentration reaches $\geq$20 nM;[70] we decreased the amount of A3C and its mutant protein to specifically test their inherent DNA binding capacity. In a titration experiment with concentrations ranging from 2 to 8 nM in steps of 2 nM of hA3C-GST and hA3C.WE-RK-GST purified protein, we detected a clear trend in the formation of DNA–protein complexes for hA3C-GST and hA3C.WE-RK-GST (Figure 9(c)). Intriguingly, DNA-protein complexes of hA3C.WE-RK-GST started appearing at the lowest protein concentration used (2 nM),

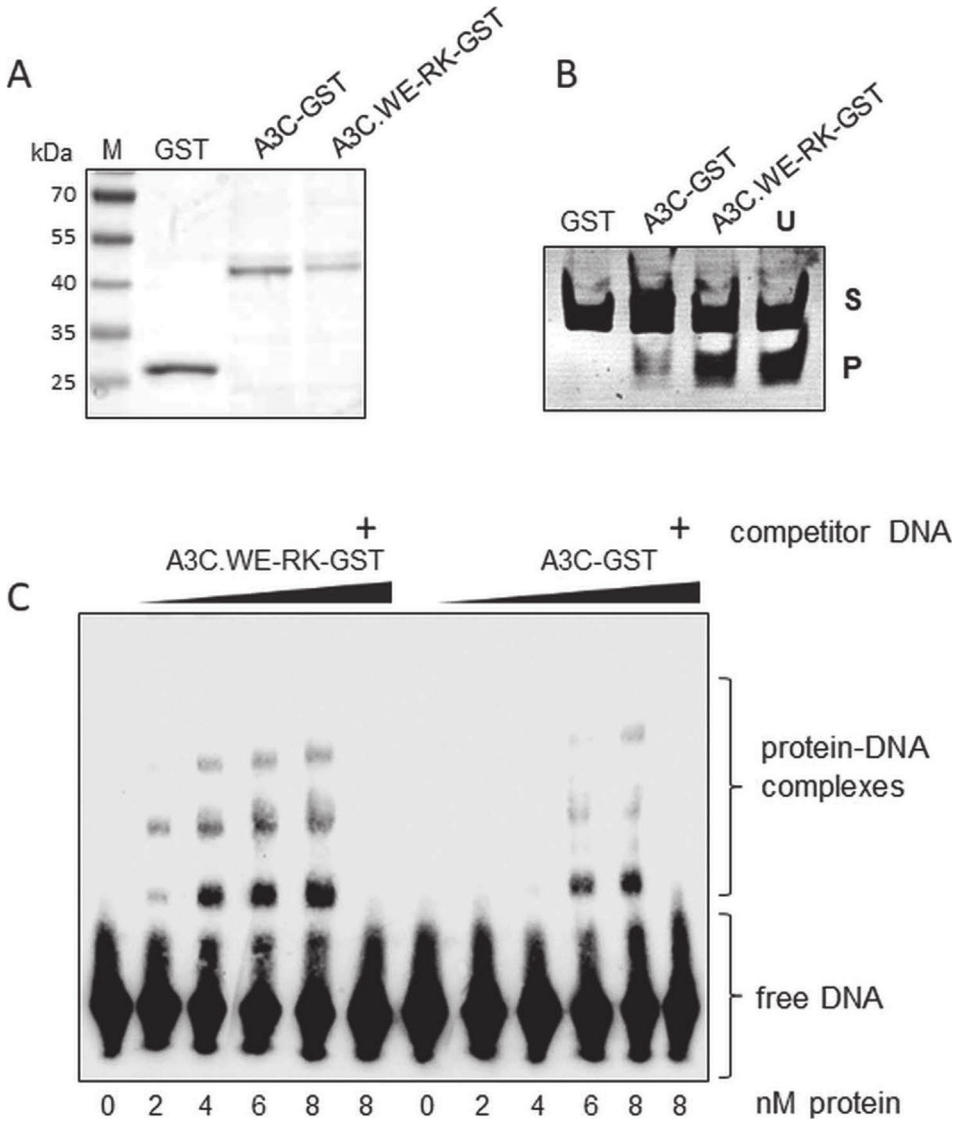


Figure 9. Recombinant hA3C.WE-RK efficiently catalyzes and displays improved interaction with ssDNA. (a) The purity of the recombinantly produced and affinity-purified proteins GST, A3C-GST, and A3C.WE-RK-GST was demonstrated by SDS-PAGE and subsequent Coomassie blue staining of the gel. The prestained protein ladder (M) indicates molecular mass. (b) *In vitro* deamination assay to examine the catalytic activity of purified GST and GST fusion proteins A3C-GST and A3C.WE-RK-GST was performed. RNAse A-treatment was included; oligonucleotide containing uracil (U) instead of cytosine served as a marker to denote the migration of the deaminated products after restriction enzyme cleavage. S-substrate, P-product. (c) EMSA with GST-tagged hA3C.WE-RK-GST and A3C-GST produced in HEK293T cells was performed with 30-nt ssDNA target DNA labelled with 3′-labelled biotin. Indicated protein concentrations (at the bottom of the blot, in nM) were titrated with 1.33 nM (20 fmol) of DNA. Presence of competitor DNA (unlabeled 80-nt DNA used in deamination assay, 200-fold molar excess added) used to demonstrate the specific binding of the protein to DNA being causative for the shift.

while hA3C-GST-DNA complexes were detected at protein concentrations $\geq$6 nM. To confirm the specificity of the DNA–protein complexes, we competed for the reaction with unlabeled DNA carrying the same nucleotide sequence as the used probe in 200-fold excess relative to that probe. The addition of the competitor DNA to the sample containing the maximum (8 nM) amount of A3C protein, efficiently disrupted the protein-DNA complex formation (Figure 9(c) and Suppl. Figure S7). Together, data from structural modeling and EMSA experiments allowed us to conclude that the two amino acid-change in loop 1 of A3C boosts the ssDNA binding capacity of A3C. Importantly, the GST moiety did not affect the binding (Suppl. Figure S7).[70]

Evolution of A3Z2 loop 1 regions in primates

Because of the strong evolutionary relationship between A3C, the CTD of A3F, and related A3Z2

proteins,[6] we performed a phylogenetic reconstruction for the A3Z2 domains in primates, using the A3Z2 sequences in the northern tree shrew as outgroup. Our analyses were performed at the A3Z2 domain level, separating the two Z2 domains of the double-domain A3D and A3F proteins, thus generating five evolutionary units: the A3D-NTD, A3F-NTD, A3C, A3D-CTD and A3F-CTD (Suppl. Figure S8). Remarkably, the results show that the A3Z2 domains underwent independent duplication in the two sister taxa, tree shrews and primates, as the three A3Z2 tree shrew sequences constitute a clear outgroup to all primate A3Z2 sequences. We identified a sharp clustering of the A3D-NTD and A3F-NTD on the one hand and of A3C, A3D-CTD, and A3F-CTD on the other hand. As to New World monkeys (Platyrrhini), we could only confidently retrieve A3C sequences from the white-faced sapajou *Cebus capucinus* and from the Ma's night monkey *Aotus nancymaae.* These sequences from A3C New World monkeys were basal to all Catarrhini (Old World monkeys and apes) A3C, A3D-CTD and A3F-CTD sequences, suggesting that the two gene duplications leading to the extant organization of A3C, A3D, and A3F occurred after the Platyrrhini/Catarrhini split 43.2 Mya (41.0–45.7 Mya) and before the Cercopithecoidea/Hominoidea (Old World monkeys/apes) split 29.44 Mya (27.95–31.35 Mya). The results show a tangled distribution within the A3D-NTD and A3F-NTD clade, and within the A3D-CTD and A3F-CTD clade. These confusing relationships are more obvious when comparing the phylogenetic reconstruction of the Z2 domains without imposing any topological constraint (Suppl. Figure S8) with a tree in which monophyly of each of the large six clades identified was enforced (Suppl. Figure S8). The tanglegram linking both, highlights those sequences whose phylogenetic position does not match the expected cluster, after the current annotation. Conversely, Catarrhini A3C sequences form a well-supported monophyletic taxon, and this A3C gene tree essentially adheres to the corresponding species tree (Figure 10). Focusing exclusively on the nodes that we could identify with confidence, we performed ancestral phylogenetic inference of the most likely amino acid sequence for the A3 loop 1 (Suppl. Figure S9) and, in parallel, performed a consensus analysis of the extant sequences (Figure 10 and Suppl. Figure S9). Our results recover

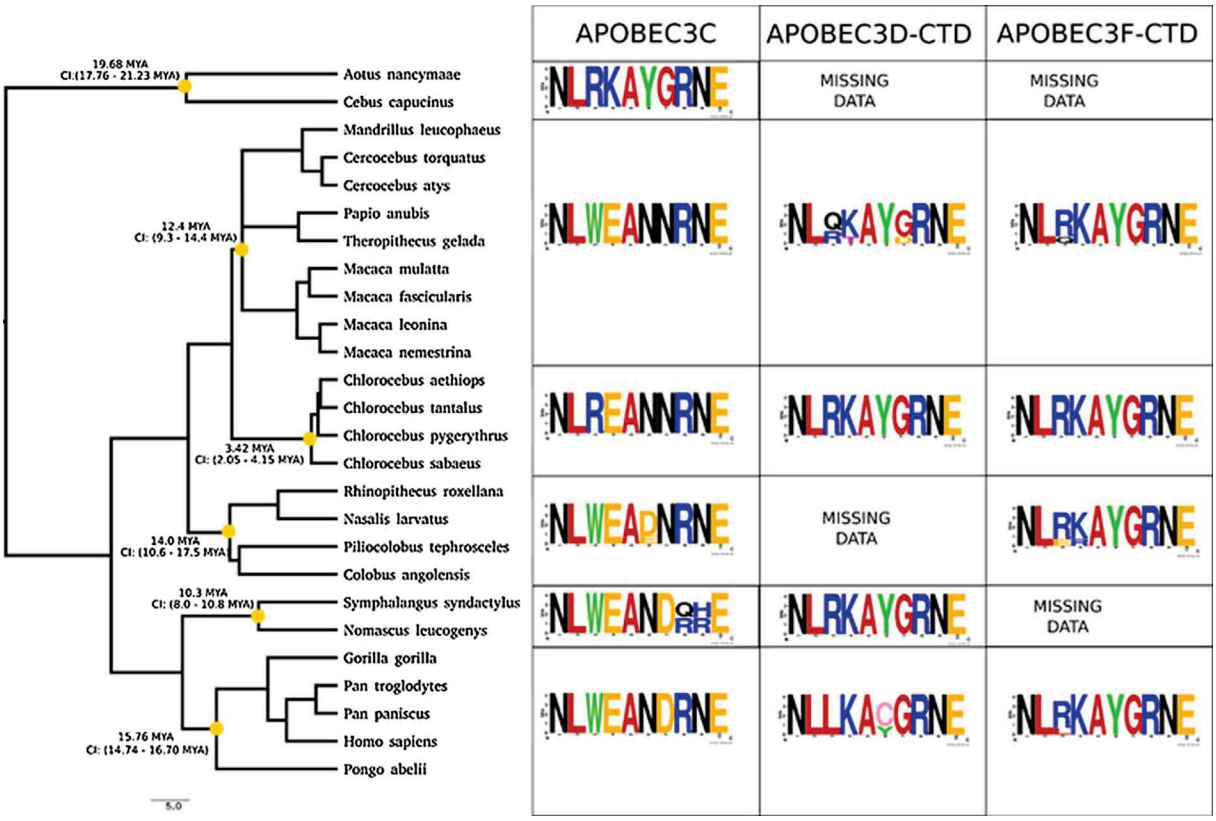


Figure 10. Species tree and one-letter amino acid sequence consensus of the loop 1 in A3C, A3D-CTD and A3F-CTD. The size of the amino acid symbol is proportional to its conservation among the sequences used. The orange dots in the species tree indicate the nodes used for consensus inference and correspond to the different rows in the table. The median values for the most recent common ancestor and the 95% confidence interval (obtained from http://www.timetree.org/) are indicated close to these reference nodes.

the well-conserved aromatic stacking stretch F[FY]FXF characteristic of all A3s. In the A3C, A3D-CTD, and A3F-CTD clade, we identified a small motif displaying striking divergent evolution flanked by conserved small hydrophobic amino acids. The most likely ancestral form is the amino acid motif LRKA, which is also the form present in extant New World monkeys A3C and the most common in extant A3F-CTD, while in the extant A3D-CTD the Arg residue is less conserved L[RLQ][KT]A (Figure 10). Strikingly, in the ancestor of Catarrhini A3C at around 29.4 Mya (27.6–31.3 Mya), this motif had already evolved to LWEA (Suppl. Figure S9), and this is the common extant form in Old World monkeys and apes (Figure 10). Only subsequently, and exclusively in the Chlorocebus lineage (African green monkeys), this change was partly reverted to LREA by a TGG > CGG transition. This reversion should have occurred after the divergence within Cercopithecinae, around 13.7 Mya (10.7–16.6 Mya) and before the speciation within Chlorocebus at 3.42 Ma (2.05–4.15 Mya) (Figure 10).

Discussion

Compared to the many studies conducted over the past decade on the HIV-1 restriction factors A3G and A3F, investigations on A3C are very limited. A small number of studies have addressed the catalytic activity and substrate binding capacity of A3C.[61,70,74,85] While the previously characterized hA3C mutants S61P and S188I boost the catalytic activity of the enzyme to a certain degree, none of these mutations is powerful enough to reduce the HIV-1Δ*vif* infectivity to the level accomplished by A3G and they do not directly partake in catalytic activity.[70,74,85] Because our repeated attempts to express A3F-CTD in human cells were not successful (Figure 1(c)),[70] we assayed A3C proteins from different Old World primate species. Due to the high level of nucleotide sequence identity between the A3C (A3Z2) paralogs (see discussion below) in the sooty mangabey monkey genome, we generated by missannotation a smmA3C-like protein with superior anti-HIV-1 and enzymatic activity. We have identified the key role of two positively-charged residues in loop 1 of this smmA3C-like protein (and of the hA3F-CTD), namely R25 and K26 in the RKYG motif. Replacing RKYG of smmA3C-like by the WEND form of this motif in hA3C abolished its anti-HIV-1 and catalytic activity. Importantly, the converse strategy of introducing the substitution WE-RK in the loop 1 of hA3C generated the potent, deaminase-dependent anti-HIV-1 enzyme hA3C.WE-RK. Consistent with these observations, our EMSA data demonstrate that residues in the loop 1 of A3C regulate protein-DNA interaction. Thus, we postulate that this more intense DNA-protein interaction is causative for the enhanced deamination activity and enhanced anti-HIV and anti-L1 activity. Similarly, Solomon and coworkers discussed that loop 1 residues of hA3G-CTD strongly interact with substrate ssDNA and that this interaction distinguishes catalytic binding from non-catalytic binding.[86] However, the loop 1 of A3 proteins likely has multiple functions, as loop 1 of A3A was found to be important for substrate specificity but not for substrate binding affinity,[87] and loop 1 of A3H, especially its residue R26, plays a triple role for RNA binding, DNA substrate recognition, and catalytic activity likely by positioning the DNA substrate in the active site for effective catalysis.[88] In accordance, our study indicates that ^{25}RK^{26} substitution in loop 1 of A3C provides the microenvironment that drives the flexibility in substrate binding and enzymatic activity.

The binding model developed here rationalizes how hA3C.RKYG can interact with the negatively charged backbone of ssDNA via the positively charged loop 1 side chains of R25 and K26 (Figure 8(c)). Like our modeling strategy, Fang *et al.* used their binding mode model of A3F-CTD with ssDNA to identify residues in the A3G-CTD important for ssDNA binding.[83] Furthermore, the increased mobility of DNA binding regions carrying the substitutions in hA3C.RKYG and hA3C.S61P.S188I, respectively, compared to hA3C (Suppl. Figure S5(b)) suggests that hA3C.RKYG and hA3C.S61P.S188I can better slide along the ssDNA than hA3C. The higher mobility of the residues may allow them to adapt more quickly to the passing ssDNA, which, together with likely stronger interactions with the ssDNA backbone, may explain the increased deaminase activity. This idea is corroborated by the MD simulations, in which the complexes including DNA loop 1 residues show more frequent interactions with the DNA in the case of hA3C.RKYG than in any of the other two variants, suggesting a stronger binding of the DNA; by contrast, in hA3C.S61P.S188I in 39.2% of the time either W25 or R30 interact with the DNA such that the DNA could be passed on from one residue to the other, assisting in the sliding-down mechanism while possibly also increasing binding affinity. In addition, loop 7 exhibits a decreased mobility in both hA3C.RKYG and hA3C.S61P.S188I compared to hA3C (Suppl. Figure S5(b)). Decreased mobility of loop 7 has been shown to predict higher deaminase activity, DNA binding, and substrate specificity of A3G and A3F, and has been reported to be also relevant for antiviral activity of A3B and A3D.[76,89–91] These structural findings can explain the differences in deaminase activity among the three variants.

Unexpectedly, our experiments also demonstrated that LINE-1 restriction by A3C, which was reported earlier to be deaminase-independent,[61] is enhanced after expression of the A3C.WE-RK variant. These data suggest that the reported RNA-dependent physical interaction

between L1 ORF1p and A3C dimers might be mediated by A3C loop 1, is partly dependent on the two amino acids W25 and E26 and is enhanced by the R25 and K26 substitutions. However, L1 inhibition by A3F was not significantly altered by the A3F.RK-WE mutations, clearly indicating that other regions (and NTD) in A3F are relevant for L1 restriction.

Because selection likely had to balance between anti-viral/anti-L1 activity and genotoxicity of A3 proteins, we wanted to characterize loop 1 residues during the evolution of the closely related A3Z2 proteins A3C, A3D CTD and A3F CTD in primates, all of them descendant of an ancestral Z2 domain that had undergone two duplication rounds.[6] In the most recent common ancestor of these enzymes that existed before the split Old World and New World primates (Catarrhini-Platyrrhini) around 43 Mya, we infer the ancestral form of the sequence of this motif in loop 1 to be LRKAYG. In New World monkeys, the A3C genes were not duplicated and are basal to the three sister clades of Catarrhini A3C, A3D-CTD, and A3F-CTD. In extant A3C sequences in New World monkeys, the loop 1 motif has notably remained unchanged and reads LRKAYG. In Catarrhini, on the contrary, the ancestral A3C sequence underwent two rapid rounds of duplication that occurred after the split with the ancestor of Platyrrhini, and before the split between the ancestors of Old World monkeys (Cercopithecoidea) and apes (Hominoidea), some 29 Mya.[6] A3F has since then been involved in an Red Queen arms race with retroviral genes.[92] In extant A3F-CTD sequences, the consensus form of the loop 1 remains LRKAYG, albeit with a certain variability of the R residue, which is exchanged with other positively charged amino acids. In extant A3D-CTD enzymes, this motif has undergone erosion, is more variable and reads L[RLQ][KT]A[YC]G. Interestingly, loop 1 in A3C experienced rapid and swift selective pressure to exchange the positively charged RK amino acids by the largely divergent chemistry of WE, yielding LWEAYG. This selective sweep occurred very rapidly, as this is the fixed form in all Catarrhini. Notoriously, and exclusively in the Chlorocebus lineage (African Green monkeys), this amino acid substitution was partly reverted to LREAYG, which is the conserved sequence in the four Chlorocebus A3C entries available (Figure 10).

Overall, our results suggest that the two duplication events that generated the extant A3C, A3D-CTD, and A3F-CTD sequences in Catarrhines released the selective pressure on two of the daughter enzymes allowing them to explore the sequence space and to evolve via sub/neofunctionalization, as proposed for Ohno's in-paralogs.[93] Thus, the A3F-CTD form of the loop 1 diverged little from the ancestral chemistry and possibly maintained the ancestral function, while the release in conservation pressure on A3D-CTD allowed the enzyme loop 1 to accumulate mutations and diverge from the ancestral state. In turn, A3C was rapidly engaged into a distinct evolutionary pathway, which is unique due to the highly divergent chemistry of loop 1 but also because A3C is the only A3Z2 monodomain enzyme of the A3 family. It must also be noted that among the descendants of the ancestral A3C in Catarrhines, only extant A3C forms a well-supported monophyletic clade (Suppl. Figs. S8 and S9). Instead, in several instances and for different species, sequences annotated in the databases as A3D-CTD clustered together with sequences annotated as A3F-CTD, and vice versa, and the same is true for the corresponding N-terminal domains (see tanglegram Suppl. Figure S8), overall resulting in a lack of support for common ancestry for the individual moieties of A3C and A3F, and preventing us from inferring the ancestral forms of the loop 1 in A3D-CTD and A3F-CTD. This lack of monophyly could simply reflect the lack of power of phylogenetic reconstruction or the potential for database misannotations when applied to genes undergoing complex evolution, including a full panel of duplications, deletions, adaptive radiation, differential selection among paralogs and Red Queen dynamics.[3,6,92,94,95] In this respect, the field is wanting for a systematisation of protocols and procedures for identifying selection signatures in genes with complex evolutionary histories.[96] This lack of resolution could also reflect a biological basis of read-through of unmatured mRNAs resulting in differentially edited or in naturally chimeric mRNAs;[9,97,98] which can hamper phylogenetic inference. Finally, the genetic architecture of the A3 locus, with the different gene copies located in tandem may favour non-homologous recombination between recently diverged, closely related sequences, and may also facilitate gene conversion between non-homologous alleles, overall leading to genetic information flow between gene copies and decoupling the true evolutionary history from our gene name and annotation-based phylogenetic reconstructions. The combined result of these novelty-generating mechanisms could be an enhanced inter-species or even inter-individual diversity in the A3 locus at either the genetic or the transcriptomic levels.[98,99] The functional impact of such gene and mRNA diversity deserves further investigation, especially in the context of personalised medicine.

In conclusion, we postulate that the loop 1 region of A3s might have a conserved role in anchoring its ssDNA substrate for efficient catalysis and that weak deamination and anti-HIV-1 activity of hA3C might have been the result of losing DNA interactions in loop 1 during its evolution. It is thus possible that genes encoding A3C proteins with loop 1 residues with a higher ssDNA affinity were too genotoxic to benefit their hosts by superior anti-viral and anti-L1 activity. Tao *et al.* noted that

the level of A3C preferentially increased upon treatment with artesunate (Art) and suggested that upregulated A3C is involved in the Art-induced DNA damage response.[100] Conceptually, we cannot rule out the possibility that the residues characterized here in loop 1 of hA3C might have an impact on recognition of unknown substrates or targets.

Materials and Methods

Cell culture

HEK293T cells (ATCC CRL-3216) were maintained in Dulbecco's high-glucose modified Eagle's medium (DMEM) (Biochrom, Berlin, Germany), supplemented with 10% fetal bovine serum (FBS), 2 mM L-glutamine, 50 units/ml penicillin, and 50 μg/ml streptomycin at 37 °C in a humidified atmosphere of 5% CO_2. Similarly, HeLa-HA cells[101] were cultured in DMEM with 10% FCS (Biowest, Nuaill , France), 2 mM L-glutamine and 20 U/ml penicillin/streptomycin (Gibco, Schwerte, Germany).

Plasmids

The HIV-1 packaging plasmid pMDLg/pRRE encodes *gag-pol*, and the pRSV-Rev for the HIV-1 *rev*.[102] The HIV-1 vector pSIN.PPT.CMV.Luc.IRES. GFP expresses the firefly luciferase and GFP.[103] HIV-1 based viral vectors were pseudotyped using the pMD.G plasmid that encodes the glycoprotein of VSV (VSV-G). SIVagm luciferase vector system was described before.[33] All A3 constructs described here were cloned in pcDNA3.1 (+) with a C-terminal hemagglutinin (HA) tag. The smmA3C-like expression plasmid was generated by exon assembly from the genomic DNA of a white-crowned mangabey (*Cercocebus torquatus lunulatus*), and the cloning strategy for smmA3C-like and the chimeras of hA3C/smmA3C-like plasmid construction was recently described.[69] The expression vector for A3G-HA was generously provided by Nathaniel R. Landau. Expression constructs hA3C, rhA3C, cpzA3C, agmA3C and A3C point mutant A3C.C97S were described before.[57,60,70]

Various point mutants hA3C.WE-RK, hA3C.ND-YG, hA3C.WE-RK.C97S, hA3C.WE-RK.S61P, hA3C.WE-RK.S61P.C97S, hA3C.WE-RK.S61P.S188I, hA3C.WE-RK.S61P.S188I.C97S, hA3F.RK-WE, smmA3C-like.E68A were generated by using site-directed mutagenesis. Similarly, single or multiple amino acid changes were made in expression vectors to produce chimera 2 mutants (C2.DH-YG, C2.RKYG-WEND, C2.R25W, C2.K26E, C2.Y28N, and C2.G29D) and smmA3C-like.RKYG-WEND. To clone C-terminal GST-tagged hA3C, hA3C.WE-RK, the ORFs were inserted between the restriction sites HindIII and XbaI in the mammalian expression construct pK-GST mammalian expression vector.[104] Individual exons of authentic smmA3C and smmA3F and smmA3F-like genes exons were amplified and cloned in pcDNA3.1. All primer sequences are listed in Suppl. Table 1.

Virus production and isolation

HEK293T cells were transiently transfected using Lipofectamine LTX and Plus reagent (Invitrogen, Karlsruhe, Germany) with an appropriate combination of HIV-1 viral vectors (600 ng pMDLg/pRRE, 600 ng pSIN.PPT.CMV.Luc.IRES. GFP, 250 ng pRSV-Rev, 150 ng pMD.G with 600 ng A3 plasmid or replaced by pcDNA3.1, unless otherwise mentioned) or SIVagm vectors (1400 ng pSIVTan-LucΔ*vif*, 150 ng pMD.G with 600 ng A3 plasmid) in 6 well plate. 48 h post-transfection, virion containing supernatants were collected and for isolation of virions, concentrated by layering on 20% sucrose cushion and centrifuged for 4 h at 14,800 rpm. Viral particles were re-suspended in mild lysis buffer (50 mM Tris (pH 8), 1 mM PMSF, 10% glycerol, 0.8% NP-40, 150 mM NaCl and 1X complete protease inhibitor).

Luciferase-based infectivity assay

HIV-1 luciferase reporter viruses were used to transduce HEK293T cells. Prior infection, the amount of reverse transcriptase (RT) in the viral particles was determined by RT assay using Cavidi HS kit Lenti RT (Cavidi Tech, Uppsala, Sweden). Normalized RT amount equivalent viral supernatants were transduced. 48 h later, luciferase activity was measured using SteadyliteHTS luciferase reagent substrate (Perkin Elmer, Rodgau, Germany) on a Berthold MicroLumat Plus luminometer (Berthold Detection Systems, Pforzheim, Germany). Transductions were done in triplicates and at least three independent experiments were performed.

Immunofluorescence microscopy

1×10^5 HeLa cells grown on polyethylene coverslips (Thermo Fisher Scientific) were co-transfected with plasmids for hemagglutinin (HA) tagged hA3C (0.25 μg) WT or hA3C.WE-RK (0.25 μg) using FuGENE transfection reagent (Promega, Wisconsin, USA). At day 2 post transfection, cells were fixed with 4% paraformaldehyde in phosphate-buffered saline (PBS) for 10 mins, permeabilized 0.1% Triton X-100 for 10 min, incubated with blocking solution (10% FBS in PBS) for 1 h, and then cells were stained with mouse anti-HA antibody (Covance, Münster, Germany) 1:1000 dilution in blocking solution for 1 h. Donkey anti-mouse Alexa Fluor 488 (Covance) was used as a secondary antibody, 1:300 dilution in blocking solution for

1 h. Finally, DAPI was used to stain nuclei for 2 minutes. The images were captured by using a 63× objective on Zeiss LSM 510 Meta laser scanning confocal microscopy (Carl Zeiss, Cologne, Germany). For the quantification of cellular localization of A3Cs, 40 randomly chosen transfected cells with A3C or A3C.WE-RK were categorized and quantified.

Immunoblot analyses

Transfected HEK293T cells were washed with PBS and lysed in radioimmunoprecipitation assay buffer (RIPA, 25 mM Tris (pH 8.0), 137 mM NaCl, 1% glycerol, 0.1% SDS, 0.5% sodium deoxycholate, 1% Nonidet P-40, 2 mM EDTA, and protease inhibitor cocktail set III [Calbiochem, Darmstadt, Germany]) 20 min on ice. Lysates were clarified by centrifugation (20 min, 14,800 rpm, 4 °C). Samples (cell/viral lysate) were boiled at 95 °C for 5 min with Roti load reducing loading buffer (Carl Roth, Karlsruhe, Germany) and subjected to SDS-PAGE followed by transfer (Semi-Dry Transfer Cell, Biorad, Munich, Germany) to a PVDF membrane (Merck Millipore, Schwalbach, Germany). Membranes were blocked with skimmed milk solution and probed with appropriate primary antibody, mouse anti-hemagglutinin (anti-HA) antibody (1:7500 dilution, MMS-101P, Covance); goat anti-GAPDH (C terminus, 1:15,000 dilution, Everest Biotech, Oxfordshire, UK); mouse anti-α-tubulin antibody (1:4000 dilution, clone B5-1-2; Sigma-Aldrich, Taufkirchen, Germany), mouse anti-capsid p24/p27 MAb AG3.0[105] (1:250 dilution, NIH AIDS Reagents); rabbit anti S6 ribosomal protein (5G10; 1:10^3 dilution in 5% BSA, Cell Signaling Technology, Leiden, The Netherlands). Secondary Abs.: anti-mouse (NA931V), anti-rabbit (NA934V) horseradish peroxidase (1:10^4 dilution, GE Healthcare) and anti-goat IgG-HRP (1:10^4 dilution, sc-2768, Santa Cruz Biotechnology, Heidelberg, Germany). Signals were visualized using ECL chemiluminescent reagent (GE Healthcare). To characterize the effect of the expression of A3 proteins and their mutants on LINE-1 (L1) reporter expression, HeLa-HA cells were lysed 48 h post-transfection using triple lysis buffer (20 mM Tris/HCl, pH 7.5; 150 mM NaCl; 10 mM EDTA; 0.1% SDS; 1% Triton X-100; 1% deoxycholate; 1x complete protease inhibitor cocktail [Roche]), clarified and 20 μg total protein were used for SDS-PAGE followed by electroblotting. HA-tagged A3 proteins and L1 ORF1p were detected using an anti-HA antibody (MMS-101P; Covance) in a 1:5000 dilution and the polyclonal rabbit-anti-L1 ORF1p antibody #984[106] in a 1:2000 dilution, respectively, in 1xPBS-T containing 5% milk powder (Suppl. Figure S4). ß-actin expression (AC-74, 1:30,000 dilution, Sigma-Aldrich Chemie GmbH) served as a loading control.

Differential DNA denaturation (3D) PCR

HEK293T cells were cultured in 6-well plates and infected with DNAse I (Thermo Fisher Scientific) treated viruses for 12 h. Cells were harvested and washed in PBS, the total DNA was isolated using DNeasy DNA isolation kit (Qiagen, Hilden, Germany). A 714-bp fragment of the luciferase gene was amplified using the primers 5′-GATATGT GGATTTCGAGTCGTC-3′ and 5′-GTCATCGTCTT TCCGTGCTC-3′. For selective amplification of the hypermutated products, the PCR denaturation temperature was lowered stepwise from 87.6 °C to 83.5 °C (83.5 °C, 84.2 °C, 85.2 °C, 86.3 °C, 87.6 °C) using a gradient thermocycler. The PCR parameters were as follows: (i) 95 °C for 5 min; (ii) 40 cycles, with 1 cycle consisting of 83.5 °C to 87.6 °C for 30 s, 55 °C for 30 s, 72 °C for 1 min; (iii) 10 min at 72 °C. PCRs were performed with Dream Taq DNA polymerase (Thermo Fisher Scientific). PCR products were stained with ethidium bromide. PCR product (smmA3C-like sample only) from the lowest denaturation ? temperature was cloned using CloneJET PCR Cloning Kit (Thermo Fisher Scientific) and sequenced. smmA3C-like protein-induced hypermutations of eleven independent clones were analysed with the Hypermut online tool (https://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html).[107] Mutated sequences (clones) carrying similar base changes were omitted and only the unique clones were presented for clarity.

In vitro DNA cytidine deamination assay

A3 proteins expressed in transfected HEK293T cells, virion-incorporated A3s, or purified GST fusion proteins were used as input. Cell lysates were prepared with mild lysis buffer 48 h post plasmid transfection. Deamination reactions were performed as described[72,108] in a 10 μL reaction volume containing 25 mM Tris pH 7.0, 2 μl of cell lysate and 100 fmol single-stranded DNA substrate (TTCA: 5′-GGATTGGTTGGTTATTTGTATAAG GAAGGTGGATTGAAGGTTCAAGAAGGTGATG GAAGTTATGTTTGGTAGATTGATGG). Samples were treated with 50 μg/ml RNAse A (Thermo Fisher Scientific). Reactions were incubated for 1 h at 37 °C and the reaction was terminated by boiling at 95 °C for 5 min. One fmol of the reaction mixture was used for PCR amplification Dream Taq polymerase (Thermo Fisher Scientific) 95 °C for 3 min, followed by 30 cycles of 61 °C for 30 s and 94 °C for 30 s using primers forward 5′-GGATTG GTTGGTTATTTGTATAAGGA and reverse 5′-CCAT CAATCTACCAAACATAACTTCCA. PCR products were digested with MseI (NEB, Frankfurt/Main, Germany), and resolved on 15% PAGE, stained with ethidium bromide (7.5 μg/ml). As a positive control, substrate oligonucleotides with TTUA instead of

TTCA were used to control the restriction enzyme digestion.[70]

L1 retrotransposition reporter assay

Relative L1 retrotransposition activity was determined by applying a rapid dual-luciferase reporter-based assay described previously.[81] Briefly, 2×10^5 HeLa-HA cells were seeded per well of a six-well plate and transfected using Fugene-HD transfection reagent (Promega) according to the manufacturer's protocol. Each well was cotransfected with 0.5 μg of the L1 retrotransposition reporter plasmid pYX017 or pYX015[81] and 0.5 μg of pcDNA3.1 or WT or mutant A3 expression construct resuspended in 3 μl Fugene-HD transfection reagent and 100 μl GlutaMAX-I-supplemented Opti-MEM I reduced-serum medium (Thermo Fisher Scientific). Three days after transfection, the medium was replaced by complete DMEM containing 2.5 μg/ml puromycin, to select for the presence of the L1 reporter plasmid harboring a puroR-expression cassette. Next day, the medium was replaced by puromycin containing DMEM medium and 48 h later, transfected cells were lysed to quantify dual-luciferase luminescence. Dual-luciferase luminescence measurement: Luminescence was measured using the Dual-Luciferase Reporter Assay System (Promega) following the manufacturer's instructions. For assays in 6-well plates, 200 μl Passive Lysis Buffer was used to lyse cells in each well; for all assays, 20 μl lysate was transferred to a solid white 96-well plate, mixed with 50 μl Luciferase Assay Reagent II, and firefly luciferase (Fluc) activity was quantified using the microplate luminometer Infinite 200PRO (Tecan, M nnedorf, Switzerland). Renilla luciferase (Rluc) activity was subsequently read after mixing 50 μl Stop & Glo Reagent into the cell lysate containing Luciferase Assay Reagent II. Data were normalized as described in the results section. L1 retrotransposition activities were expressed as normalized luminescence ratios (NLR) relative to the background signal obtained after cotransfection of pcDNA3.1(+) and pYX015 coding for the retrotransposition-defective L1RP/JM111 element. The NLR resulting from cotransfection of pYX015 and pcDNA3.1(+) was set as 1.

Protein sequence alignment and visualization

Sequence alignment of hA3C and smmA3C-like protein was done by using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). The alignment file was then submitted to ESPript 3.0[109] (espript.ibcp.fr) to calculate the similarity and identity of residues between both proteins and to represent the pairwise sequence alignment. Cartoon model of the crystal structure of A3C (PDB 3VOW) was constructed using PyMOL (PyMOL Molecular Graphics System version 1.5.0.4; Schr dinger, Portland, OR).

Structural model building of protein-DNA complexes

The structural models of hA3C or hA3C.RKYG binding to ssDNA were generated by first aligning the X-ray crystal structure of rhA3G-NTD (PDB ID 5K82[82]) onto the X-ray crystal structure of hA3F-CTD (PDB ID 5W2M[83]), the latter of which was co-crystallized with ssDNA. Subsequently, the hA3C X-ray crystal structure (PDB ID 3VOW[68]) was aligned onto the NTD of rhA3G, which is structurally similar to hA3C. The ssDNA and the interface region of hA3C were subsequently relaxed in the presence of each other using Maestro.[110] The same program was used to mutate hA3C to obtain the hA3C.RKYG and hA3C.S61P.S188I variants, which were again relaxed in the presence of the ssDNA. Similarly, we obtained hA3C, hA3C.RKYG, and hA3C.S61P.S188I ssDNA binding models based on the ssDNA-binding X-ray crystal structure of hA3A (PDB ID 5SWW[111]), a much relevant model similar to 6BUX.[112] These three DNA complex structures were later used for MD simulations as they include a cytidine residue in the active center.

Molecular dynamics simulations

hA3C, hA3C.RKYG, and hA3C.S61P.S188I were subjected to MD simulations. For this, the above-mentioned structures without the DNA were N- and C-terminally capped with ACE and NME, respectively. The three variants were protonated with PROPKA[113] according to pH 7.4, neutralized by adding counter ions, and solvated in an octahedral box of TIP3P water[114] with a minimal water shell of 12 Å around the solute. The Amber package of molecular simulation software[115] and the ff14SB force field[116] were used to perform the MD simulations. For the Zn^{2+}-ions the Li-Merz parameters for two-fold positively charged metal ions[117] were used. To cope with long-range interactions, the "Particle Mesh Ewald" method[118] was used; the SHAKE algorithm[119] was applied to bonds involving hydrogen atoms. As hydrogen mass repartitioning[120] was utilized, the time step for all MD simulations was 4 fs with a direct-space, non-bonded cut-off of 8 Å.

In the beginning, 17,500 steps of steepest descent and conjugate gradient minimization were performed; during 2500, 10000, and 5000 steps positional harmonic restraints with force constants of 25 kcal mol^{-1} $Å^{-2}$, 5 kcal mol^{-1} $Å^{-2}$, and zero, respectively, were applied to the solute atoms. Thereafter, 50 ps of NVT (constant number of particles, volume, and temperature) MD simulations were conducted to heat up the system to 100 K, followed by 300 ps of NPT (constant number of particles, pressure, and temperature) MD simulations to adjust the density of the simulation box to a pressure of 1 atm and to heat the system to 300 K. During these steps, a harmonic potential with a force constant of

10 kcal mol^{-1} $Å^{-2}$ was applied to the solute atoms. As the final step in thermalization, 300 ps of NVT-MD simulations were performed while gradually reducing the restraint forces on the solute atoms to zero within the first 100 ps of this step. Afterwards, five independent production runs of NVT-MD simulations with 2 μs length each were performed. For this purpose, the starting temperatures of the MD simulations at the beginning of the thermalization were varied by a fraction of one Kelvin. MD simulation of those three variants in complex with ssDNA were performed similarly, treating the DNA with the OL15 force field[121] and performing ten independent production runs of NVT-MD simulations with 2 μs length each. To evaluate the interactions between loop 1 (residues 25–30) of the three variants and the ssDNA present in the complexes, we employed two different measures using CPPTRAJ[122]. First, we used the h-bond command to detect hydrogen bonds between residues in loop 1 and the ssDNA. Second, we measured the minimal distance of the side chain atoms, not including C_β of the respective residues, and the DNA for each snapshot of the MD simulations and correlated both (Suppl. Figure S6), considering a larger distance cut-off of 4 Å to detect interactions between the side chains and DNA. The minimal distance over time for residue 30 can be seen in Suppl. Figure S10 (a)–(c) and the root mean square deviation (RMSD) over time is shown in Suppl. Figure S10(d)–(f). The latter figure indicates that the systems structurally stabilized after ~250 ns.

Expression and purification of recombinant GST-tagged hA3C and hA3C.WE-RK from HEK293T cells

Recombinant C-terminal GST-tagged hA3C and hA3C.WE-RK were expressed in HEK293T cells and purified by affinity chromatography using Glutathione Sepharose 4B beads (GE Healthcare) as described previously.[70] Cells were lysed 48 h later with mild lysis buffer [50 mM Tris (pH 8), 1 mM PMSF, 10% glycerol, 0.8% NP-40, 150 mM NaCl, and 1X complete protease inhibitor and incubated with GST beads. After 2 h incubation at 4 °C in end-over-end rotation, GST beads were washed twice with wash buffer containing 50 mM Tris (pH 8.0), 5 mM 2-ME, 10% glycerol and 500 mM NaCl. The bound GST hA3C and hA3C.WE-RK proteins were eluted with wash buffer containing 20 mM reduced glutathione. The proteins were 90–95% pure as checked on 15% SDS-PAGE followed by Coomassie blue staining. Protein concentrations were estimated by Bradford's method.

Electrophoretic mobility shift assay (EMSA) with hA3C-GST and hA3C.WE-RK-GST

EMSA was performed as described previously.[70,84,123] We mixed 1.33 nM (20 fmol) of 3′ biotinylated DNA (30-TTC-Bio-TEG purchased from Eurofins Genomics, Ebersberg Germany) with 10 mM Tris (pH – 7.5), 100 mM KCl, 10 mM MgCl2, 1 mM DTT, 2% glycerol, and the respective amount of recombinant proteins in a 15 μl reaction mixture, and incubated at room temperature for 30 min. The reaction mixture containing the protein-DNA complexes were resolved on a 5% native PAGE gel on ice and transferred to a nylon membrane (Amersham Hybond-XL, GE healthcare) using 0.5 X TBE. After the transfer, the membrane containing protein-DNA complexes were cross-linked by UV radiation with 312-nm bulb for 15 min. Chemiluminescent detection of biotinylated DNA was carried out according to the manufacturer's instruction (Thermo Scientific, LightShift Chemiluminescence EMSA Kit).

Phylogenetic inference

In order to study the evolution of the A3-Z2 domains, a representative set of 61 primate A3C, A3D, and A3F gene sequences were collected from GenBank (https://www.ncbi.nlm.nih.gov/genbank), as follows: 26 A3C sequences, 12 A3D sequences, and 21 A3F sequences (full list available in Suppl. Table 2). The phylogenetic relationships and divergence times among the species used were retrieved from http://www.timetree.org (Suppl. Figure S8). A3 sequences from the northern tree shrew *Tupaia belangeri* were included as an outgroup to the primate ones. As A3D and A3F sequences contain each two Z2 domains, they were split into the corresponding N and C termini. The alignments were performed at the amino acid level using MAFFTv7.380 (http://mafft.cbrc.jp/alignment/software/).[124] Phylogenetic inference was performed using RAxMLv8;[125] at either the nucleotide level under the GTR+Γ model or at the amino acid level under the LG+Γ model. Node support was evaluated applying 5000 bootstrap cycles. Additionally, phylogenies at the nucleotide level were also calculated after introducing constraints in the tree, forcing monophyly of each clade A3D_N and C termini, A3F_N and C termini, New World monkeys A3C, and catarrhine A3C. Differences in maximum likelihood between alternative topologies for the same alignment were evaluated by the Shimodaira-Hasegawa test. Ancestral state reconstruction of amino acids in the loop A3-Z2 loop 1 was performed only for the supported clades using RAxMLv8. A tanglegram with the two phylogenies was drawn with Dendroscope v3.6.3.[126] Final layouts were done with Inkscape 0.92.4.

Statistical analysis

Data were represented as the mean with SD in all bar diagrams. Statistically significant differences between two groups were analyzed using the unpaired Student's *t*-test with GraphPad Prism version 5 (GraphPad Software, San Diego, CA,

USA). A minimum *p*-value of 0.05 was considered as statistically significant.

CRediT authorship contribution statement

Ananda Ayyappan Jaguva Vasudevan: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Kannan Balakrishnan:** Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing - review & editing. **Christoph G.W. Gertzen:** Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing - review & editing. **Fanni Borvető:** Data curation, Formal analysis, Investigation, Visualization, Writing - review & editing. **Zeli Zhang:** Formal analysis, Writing - review & editing. **Anucha Sangwiman:** Formal analysis, Investigation, Writing - review & editing. **Ulrike Held:** Data curation, Formal analysis, Investigation, Writing - review & editing. **Caroline Küstermann:** Data curation, Formal analysis, Investigation, Visualization, Writing - review & editing. **Sharmistha Banerjee:** Data curation, Formal analysis, Methodology, Writing - review & editing. **Gerald G. Schumann:** Data curation, Formal analysis, Resources, Visualization, Writing - review & editing. **Dieter Häussinger:** Data curation, Formal analysis, Supervision, Writing - review & editing. **Ignacio G. Bravo:** Data curation, Formal analysis, Investigation, Resources, Visualization, Writing - review & editing. **Holger Gohlke:** Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Visualization, Writing - review & editing. **Carsten Münk:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing - review & editing.

Acknowledgments

We thank Wioletta Hörschken for excellent technical assistance, Boris Görg for microscopy support, and Wolfgang A. Schulz for kindly proofreading the manuscript. We thank Alejandro Moisés Barbero Amézaga (University Francisco de Vitoria/Faculty of Experimental Sciences, Madrid, Spain) for his excellent technical support, and the University Francisco de Vitoria/Faculty of Experimental Sciences for financial support of A. M.B.A. We thank Michael Emerman, Jens-Ove Heckel, Henning Hofmann, Yasumasa Iwatani, Nathanial R. Landau, Neeltje Kootstra, Bryan Cullen, Jonathan Stoye, Harald Wodrich, and Jörg Zielonka for reagents. The following reagents were obtained through the NIH AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH: a monoclonal antibody to HIV-1 p24 (AG3.0) from Jonathan Allan. FB and IGB acknowledge the IRD itrop HPC (South Green Platform) at IRD Montpellier for providing computing resources. HG is grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich-Heine-University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing (NIC) to HG on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user ID: HKF7). Graphical abstract was designed with BioRender.com.

Funding

This work was supported by a grant from the research commission of the medical faculty of the Heinrich-Heine-University Düsseldorf (grant #2019-13 to CM and HG). KB is supported by the German Academic Exchange Service (DAAD). ZZ was supported by China Scholarship Council (CSC). CK and GGS are supported by the German Ministry of Health (grant # G115F020001). CM is supported by the Heinz-Ansmann Foundation for AIDS Research. The Center for Structural Studies is funded by the Deutsche Forschungsgemeinschaft (DFG Grant number 417919780 and INST 208/761-1 FUGG).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2020.10.014.

Received 31 May 2020;
Accepted 9 October 2020;
Available online 15 October 2020

Keywords:
APOBEC3C_A3F_cytidine deaminase;
sooty mangabey monkey;
human immunodeficiency virus (HIV);
LINE-1;
evolution

† A.A.J.V and K.B contributed equally to this article.
‡ La Jolla Institute for Immunology, La Jolla, CA 92037, USA.

References

1. Goila-Gaur, R., Strebel, K., (2008). HIV-1 Vif, APOBEC, and intrinsic immunity. *Retrovirology*, **5**, 51.
2. Harris, R.S., Dudley, J.P., (2015). APOBECs and virus restriction. *Virology*, **479–480**, 131–145.
3. Salter, J.D., Bennett, R.P., Smith, H.C., (2016). The APOBEC protein family: united by structure, divergent in function. *Trends Biochem. Sci.*, **41**, 578–594.
4. Silvas, T.V., Schiffer, C.A., (2019). APOBEC3s: DNA-editing human cytidine deaminases. *Protein Sci.: Publ. Protein Soc.*, **28**, 1552–1566.
5. Jarmuz, A., Chester, A., Bayliss, J., Gisbourne, J., Dunham, I., Scott, J., et al., (2002). An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics*, **79**, 285–296.
6. Münk, C., Willemsen, A., Bravo, I.G., (2012). An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. *BMC Evol. Biol.*, **12**, 71.
7. LaRue, R.S., Jonsson, S.R., Silverstein, K.A., Lajoie, M., Bertrand, D., El-Mabrouk, N., et al., (2008). The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals. *BMC Mol. Biol.*, **9**, 104.
8. LaRue, R.S., Andresdottir, V., Blanchard, Y., Conticello, S.G., Derse, D., Emerman, M., et al., (2009). Guidelines for naming nonprimate APOBEC3 genes and proteins. *J. Virol.*, **83**, 494–497.
9. Münk, C., Beck, T., Zielonka, J., Hotz-Wagenblatt, A., Chareza, S., Battenberg, M., et al., (2008). Functions, structure, and read-through alternative splicing of feline APOBEC3 genes. *Genome Biol.*, **9**, R48.
10. Sheehy, A.M., Gaddis, N.C., Choi, J.D., Malim, M.H., (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, **418**, 646–650.
11. Zhang, H., Yang, B., Pomerantz, R.J., Zhang, C., Arunachalam, S.C., Gao, L., (2003). The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature*, **424**, 94–98.
12. Bishop, K.N., Holmes, R.K., Sheehy, A.M., Davidson, N. O., Cho, S.J., Malim, M.H., (2004). Cytidine deamination of retroviral DNA by diverse APOBEC proteins. *Curr. Biol.*, **14**, 1392–1396.
13. Vasudevan, A.A., Smits, S.H., Hoppner, A., Häussinger, D., Koenig, B.W., Münk, C., (2013). Structural features of antiviral DNA cytidine deaminases. *Biol. Chem.*, **394**, 1357–1370.
14. Zennou, V., Perez-Caballero, D., Gottlinger, H., Bieniasz, P.D., (2004). APOBEC3G incorporation into human immunodeficiency virus type 1 particles. *J. Virol.*, **78**, 12058–12061.
15. Luo, K., Liu, B., Xiao, Z., Yu, Y., Yu, X., Gorelick, R., et al., (2004). Amino-terminal region of the human immunodeficiency virus type 1 nucleocapsid is required for human APOBEC3G packaging. *J. Virol.*, **78**, 11841–11852.
16. Svarovskaia, E.S., Xu, H., Mbisa, J.L., Barr, R., Gorelick, R.J., Ono, A., et al., (2004). Human apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like 3G (APOBEC3G) is incorporated into HIV-1 virions through interactions with viral and nonviral RNAs. *J. Biol. Chem.*, **279**, 35822–35828.
17. Huthoff, H., Malim, M.H., (2007). Identification of amino acid residues in APOBEC3G required for regulation by human immunodeficiency virus type 1 Vif and Virion encapsidation. *J. Virol.*, **81**, 3807–3815.
18. Schafer, A., Bogerd, H.P., Cullen, B.R., (2004). Specific packaging of APOBEC3G into HIV-1 virions is mediated by the nucleocapsid domain of the gag polyprotein precursor. *Virology*, **328**, 163–168.
19. Burnett, A., Spearman, P., (2007). APOBEC3G multimers are recruited to the plasma membrane for packaging into human immunodeficiency virus type 1 virus-like particles in an RNA-dependent process requiring the NC basic linker. *J. Virol.*, **81**, 5000–5013.
20. Browne, E.P., Allers, C., Landau, N.R., (2009). Restriction of HIV-1 by APOBEC3G is cytidine deaminase-dependent. *Virology*, **387**, 313–321.
21. Harris, R.S., Bishop, K.N., Sheehy, A.M., Craig, H.M., Petersen-Mahrt, S.K., Watt, I.N., et al., (2003). DNA deamination mediates innate immunity to retroviral infection. *Cell*, **113**, 803–809.
22. Yu, Q., König, R., Pillai, S., Chiles, K., Kearney, M., Palmer, S., et al., (2004). Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nature Struct. Mol. Biol.*, **11**, 435–442.
23. Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., Trono, D., (2003). Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature*, **424**, 99–103.
24. Iwatani, Y., Chan, D.S., Wang, F., Maynard, K.S., Sugiura, W., Gronenborn, A.M., et al., (2007). Deaminase-independent inhibition of HIV-1 reverse transcription by APOBEC3G. *Nucleic Acids Res.*, **35**, 7096–7108.
25. Holmes, R.K., Koning, F.A., Bishop, K.N., Malim, M.H., (2007). APOBEC3F can inhibit the accumulation of HIV-1 reverse transcription products in the absence of hypermutation. Comparisons with APOBEC3G. *J. Biol. Chem.*, **282**, 2587–2595.
26. Münk, C., Jensen, B.E., Zielonka, J., Häussinger, D., Kamp, C., (2012). Running loose or getting lost: How HIV-1 counters and capitalizes on APOBEC3-induced mutagenesis through Its Vif protein. *Viruses.*, **4**, 3132–3161.
27. Bishop, K.N., Holmes, R.K., Malim, M.H., (2006). Antiviral potency of APOBEC proteins does not correlate with cytidine deamination. *J. Virol.*, **80**, 8450–8458.
28. Mbisa, J.L., Bu, W., Pathak, V.K., (2010). APOBEC3F and APOBEC3G inhibit HIV-1 DNA integration by different mechanisms. *J. Virol.*, **84**, 5250–5259.
29. Strebel, K., (2005). APOBEC3G & HTLV-1: inhibition without deamination. *Retrovirology*, **2**, 37.
30. Mehle, A., Strack, B., Ancuta, P., Zhang, C., McPike, M., Gabuzda, D., (2004). Vif overcomes the innate antiviral activity of APOBEC3G by promoting its degradation in the ubiquitin-proteasome pathway. *J. Biol. Chem.*, **279**, 7792–7798.
31. Sheehy, A.M., Gaddis, N.C., Malim, M.H., (2003). The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nature Med.*, **9**, 1404–1407.
32. Yu, X., Yu, Y., Liu, B., Luo, K., Kong, W., Mao, P., et al., (2003). Induction of APOBEC3G ubiquitination and

degradation by an HIV-1 Vif-Cul5-SCF complex. *Science*, **302**, 1056–1060.
33. Mariani, R., Chen, D., Schrofelbauer, B., Navarro, F., König, R., Bollman, B., et al., (2003). Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell*, **114**, 21–31.
34. Bogerd, H.P., Doehle, B.P., Wiegand, H.L., Cullen, B.R., (2004). A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 3770–3774.
35. Mangeat, B., Turelli, P., Liao, S., Trono, D., (2004). A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action. *J. Biol. Chem.*, **279**, 14481–14483.
36. Zhang, W., Huang, M., Wang, T., Tan, L., Tian, C., Yu, X., et al., (2008). Conserved and non-conserved features of HIV-1 and SIVagm Vif mediated suppression of APOBEC3 cytidine deaminases. *Cell. Microbiol.*, **10**, 1662–1675.
37. Smith, J.L., Pathak, V.K., (2010). Identification of specific determinants of human APOBEC3F, APOBEC3C, and APOBEC3DE and African green monkey APOBEC3F that interact with HIV-1 Vif. *J. Virol.*, **84**, 12599–12608.
38. Zhang, Z., Gu, Q., de Manuel, Montero M, Bravo, I.G., Marques-Bonet, T., Häussinger, D., et al., (2017). Stably expressed APOBEC3H forms a barrier for cross-species transmission of simian immunodeficiency virus of chimpanzee to humans. *PLoS Pathog.*, **13**, e1006746.
39. Dang, Y., Wang, X., Esselman, W.J., Zheng, Y.H., (2006). Identification of APOBEC3DE as another antiretroviral factor from the human APOBEC family. *J. Virol.*, **80**, 10522–10533.
40. Wiegand, H.L., Doehle, B.P., Bogerd, H.P., Cullen, B.R., (2004). A second human antiretroviral factor, APOBEC3F, is suppressed by the HIV-1 and HIV-2 Vif proteins. *EMBO J.*, **23**, 2451–2458.
41. Zheng, Y.H., Irwin, D., Kurosu, T., Tokunaga, K., Sata, T., Peterlin, B.M., (2004). Human APOBEC3F is another host factor that blocks human immunodeficiency virus type 1 replication. *J. Virol.*, **78**, 6073–6076.
42. Hultquist, J.F., Lengyel, J.A., Refsland, E.W., LaRue, R. S., Lackey, L., Brown, W.L., et al., (2011). Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. *J. Virol.*, **85**, 11220–11234.
43. Burns, M.B., Lackey, L., Carpenter, M.A., Rathore, A., Land, A.M., Leonard, B., et al., (2013). APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*, **494**, 366–370.
44. Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S. A., Fargo, D., Stojanov, P., et al., (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genet.*, **45**, 970–976.
45. Buisson, R., Langenbucher, A., Bowen, D., Kwan, E.E., Benes, C.H., Zou, L., et al., (2019). Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science*, **364**
46. Cortez, L.M., Brown, A.L., Dennis, M.A., Collins, C.D., Brown, A.J., Mitchell, D., et al., (2019). APOBEC3A is a prominent cytidine deaminase in breast cancer. *PLoS Genet.*, **15**, e1008545.
47. Henderson, S., Fenton, T., (2015). APOBEC3 genes: retroviral restriction factors to cancer drivers. *Trends Mol. Med.*, **21**, 274–284.
48. Swanton, C., McGranahan, N., Starrett, G.J., Harris, R.S., (2015). APOBEC enzymes: Mutagenic fuel for cancer evolution and heterogeneity. *Cancer Discov.*, **5**, 704–712.
49. Green, A.M., Weitzman, M.D., (2019). The spectrum of APOBEC3 activity: From anti-viral agents to anti-cancer opportunities. *DNA Repair*, **83**, 102700.
50. Olson, M.E., Harris, R.S., Harki, D.A., (2018). APOBEC enzymes as targets for virus and cancer therapy. *Cell Chemical Biol.*, **25**, 36–49.
51. Muckenfuss, H., Hamdorf, M., Held, U., Perkovic, M., Lower, J., Cichutek, K., et al., (2006). APOBEC3 proteins inhibit human LINE-1 retrotransposition. *J. Biol. Chem.*, **281**, 22161–22172.
52. Yu, Q., Chen, D., König, R., Mariani, R., Unutmaz, D., Landau, N.R., (2004). APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. *J. Biol. Chem.*, **279**, 53379–53386.
53. Langlois, M.A., Beale, R.C., Conticello, S.G., Neuberger, M.S., (2005). Mutational comparison of the single-domained APOBEC3C and double-domained APOBEC3F/G anti-retroviral cytidine deaminases provides insight into their DNA target site specificities. *Nucleic Acids Res.*, **33**, 1913–1923.
54. Suspene, R., Guetard, D., Henry, M., Sommer, P., Wain-Hobson, S., Vartanian, J.P., (2005). Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 8321–8326.
55. Baumert, T.F., Rosler, C., Malim, M.H., von Weizsacker, F., (2007). Hepatitis B virus DNA is subject to extensive editing by the human deaminase APOBEC3C. *Hepatology*, **46**, 682–689.
56. Vartanian, J.P., Guetard, D., Henry, M., Wain-Hobson, S., (2008). Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. *Science*, **320**, 230–233.
57. Stauch, B., Hofmann, H., Perkovic, M., Weisel, M., Kopietz, F., Cichutek, K., et al., (2009). Model structure of APOBEC3C reveals a binding pocket modulating ribonucleic acid interaction required for encapsidation. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 12079–12084.
58. Ahasan, M.M., Wakae, K., Wang, Z., Kitamura, K., Liu, G., Koura, M., et al., (2015). APOBEC3A and 3C decrease human papillomavirus 16 pseudovirion infectivity. *Biochem. Biophys. Res. Commun.*, **457**, 295–299.
59. Suspene, R., Aynaud, M.M., Koch, S., Pasdeloup, D., Labetoulle, M., Gaertner, B., et al., (2011). Genetic editing of herpes simplex virus 1 and Epstein-Barr herpesvirus genomes by human APOBEC3 cytidine deaminases in culture and in vivo. *J. Virol.*, **85**, 7594–7602.
60. Perkovic, M., Schmidt, S., Marino, D., Russell, R.A., Stauch, B., Hofmann, H., et al., (2009). Species-specific inhibition of APOBEC3C by the prototype foamy virus protein bet. *J. Biol. Chem.*, **284**, 5819–5826.
61. Horn, A.V., Klawitter, S., Held, U., Berger, A., Vasudevan, A.A., Bock, A., et al., (2014). Human LINE-1 restriction by APOBEC3C is deaminase independent and mediated by an ORF1p interaction that affects LINE reverse transcriptase activity. *Nucleic Acids Res.*, **42**, 396–416.
62. Hultquist, J.F., Binka, M., LaRue, R.S., Simon, V., Harris, R.S., (2012). Vif proteins of human and simian immunodeficiency viruses require cellular CBFbeta to degrade APOBEC3 restriction factors. *J. Virol.*, **86**, 2874–2877.

63. Bonvin, M., Achermann, F., Greeve, I., Stroka, D., Keogh, A., Inderbitzin, D., et al., (2006). Interferon-inducible expression of APOBEC3 editing enzymes in human hepatocytes and inhibition of hepatitis B virus replication. *Hepatology*, **43**, 1364–1374.
64. Refsland, E.W., Hultquist, J.F., Harris, R.S., (2012). Endogenous origins of HIV-1 G-to-A hypermutation and restriction in the nonpermissive T cell line CEM2n. *PLoS Pathog.*, **8**, e1002800.
65. Bourara, K., Liegler, T.J., Grant, R.M., (2007). Target cell APOBEC3C can induce limited G-to-A mutation in HIV-1. *PLoS Pathog.*, **3**, 1477–1485.
66. Refsland, E.W., Stenglein, M.D., Shindo, K., Albin, J.S., Brown, W.L., Harris, R.S., (2010). Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucleic Acids Res.*, **38**, 4274–4284.
67. Abdel-Mohsen, M., Raposo, R.A., Deng, X., Li, M., Liegler, T., Sinclair, E., et al., (2013). Expression profile of host restriction factors in HIV-1 elite controllers. *Retrovirology*, **10**, 106.
68. Kitamura, S., Ode, H., Nakashima, M., Imahashi, M., Naganawa, Y., Kurosawa, T., et al., (2012). The APOBEC3C crystal structure and the interface for HIV-1 Vif binding. *Nature Struct. Mol. Biol.*, **19**, 1005–1010.
69. Zhang, Z., Gu, Q., Jaguva Vasudevan, A.A., Jeyaraj, M., Schmidt, S., Zielonka, J., et al., (2016). Vif Proteins from Diverse Human Immunodeficiency Virus/Simian Immunodeficiency Virus Lineages Have Distinct Binding Sites in A3C. *J. Virol.*, **90**, 10193–10208.
70. Jaguva Vasudevan, A.A., Hofmann, H., Willbold, D., Häussinger, D., Koenig, B.W., Münk, C., (2017). Enhancing the catalytic deamination activity of APOBEC3C is insufficient to inhibit Vif-deficient HIV-1. *J. Mol. Biol.*, **429**, 1171–1191.
71. Suspene, R., Henry, M., Guillot, S., Wain-Hobson, S., Vartanian, J.P., (2005). Recovery of APOBEC3-edited human immunodeficiency virus G->A hypermutants by differential DNA denaturation PCR. *J. Gen. Virol.*, **86**, 125–129.
72. Nowarski, R., Britan-Rosich, E., Shiloach, T., Kotler, M., (2008). Hypermutation by intersegmental transfer of APOBEC3G cytidine deaminase. *Nature Struct. Mol. Biol.*, **15**, 1059–1066.
73. Jaguva Vasudevan, A.A., Kreimer, U., Schulz, W.A., Krikoni, A., Schumann, G.G., Häussinger, D., et al., (2018). APOBEC3B activity is prevalent in urothelial carcinoma cells and only slightly affected by LINE-1 expression. *Front. Microbiol.*, **9**, 2088.
74. Wittkopp, C.J., Adolph, M.B., Wu, L.I., Chelico, L., Emerman, M., (2016). A single nucleotide polymorphism in human APOBEC3C enhances restriction of lentiviruses. *PLoS Pathog.*, **12**, e1005865.
75. Hache, G., Liddament, M.T., Harris, R.S., (2005). The retroviral hypermutation specificity of APOBEC3F and APOBEC3G is governed by the C-terminal DNA cytosine deaminase domain. *J. Biol. Chem.*, **280**, 10920–10924.
76. Chen, Q., Xiao, X., Wolfe, A., Chen, X.S., (2016). The in vitro biochemical characterization of an HIV-1 restriction factor APOBEC3F: Importance of loop 7 on both CD1 and CD2 for DNA binding and deamination. *J. Mol. Biol.*, **428**, 2661–2670.
77. Wan, L., Nagata, T., Katahira, M., (2018). Influence of the DNA sequence/length and pH on deaminase activity, as well as the roles of the amino acid residues around the catalytic center of APOBEC3F. *Phys. Chem. Chem. Phys.*, **20**, 3109–3117.
78. Nakashima, M., Ode, H., Kawamura, T., Kitamura, S., Naganawa, Y., Awazu, H., et al., (2016). Structural insights into HIV-1 Vif-APOBEC3F interaction. *J. Virol.*, **90**, 1034–1047.
79. Schumann, G.G., (2007). APOBEC3 proteins: major players in intracellular defence against LINE-1-mediated retrotransposition. *Biochem. Soc. Trans.*, **35**, 637–642.
80. Schumann, G.G., Gogvadze, E.V., Osanai-Futahashi, M., Kuroki, A., Münk, C., Fujiwara, H., et al., (2010). Unique functions of repetitive transcriptomes. *Intl. Rev. Cell Mol. Biol.*, **285**, 115–188.
81. Xie, Y., Rosser, J.M., Thompson, T.L., Boeke, J.D., An, W., (2011). Characterization of L1 retrotransposition with high-throughput dual-luciferase assays. *Nucleic Acids Res.*, **39**, e16.
82. Xiao, X., Li, S.X., Yang, H., Chen, X.S., (2016). Crystal structures of APOBEC3G N-domain alone and its complex with DNA. *Nature Commun.*, **7**, 12193.
83. Fang, Y., Xiao, X., Li, S.X., Wolfe, A., Chen, X.S., (2018). Molecular interactions of a DNA modifying enzyme APOBEC3F catalytic domain with a single-stranded DNA. *J. Mol. Biol.*, **430**, 87–101.
84. Marino, D., Perkovic, M., Hain, A., Jaguva Vasudevan, A. A., Hofmann, H., Hanschmann, K.M., et al., (2016). APOBEC4 enhances the replication of HIV-1. *PLoS ONE*, **11**, e0155422.
85. Adolph, M.B., Ara, A., Feng, Y., Wittkopp, C.J., Emerman, M., Fraser, J.S., et al., (2017). Cytidine deaminase efficiency of the lentiviral viral restriction factor APOBEC3C correlates with dimerization. *Nucleic Acids Res.*, **45**, 3378–3394.
86. Solomon, W.C., Myint, W., Hou, S., Kanai, T., Tripathi, R., Kurt Yilmaz, N., et al., (2019). Mechanism for APOBEC3G catalytic exclusion of RNA and non-substrate DNA. *Nucleic Acids Res.*, **47**, 7676–7689.
87. Ziegler, S.J., Hu, Y., Devarkar, S.C., Xiong, Y., (2019). APOBEC3A loop 1 is a determinant for ssDNA binding and deamination. *Biochemistry*,.
88. Bohn, J.A., DaSilva, J., Kharytonchyk, S., Mercedes, M., Vosters, J., Telesnitsky, A., et al., (2019). Flexibility in nucleic acid binding is central to APOBEC3H antiviral activity. *J. Virol.*,.
89. Rathore, A., Carpenter, M.A., Demir, O., Ikeda, T., Li, M., Shaban, N.M., et al., (2013). The local dinucleotide preference of APOBEC3G can be altered from 5'-CC to 5'-TC by a single amino acid substitution. *J. Mol. Biol.*, **425**, 4442–4454.
90. Siu, K.K., Sultana, A., Azimi, F.C., Lee, J.E., (2013). Structural determinants of HIV-1 Vif susceptibility and DNA binding in APOBEC3F. *Nature Commun.*, **4**, 2593.
91. Dang, Y., Abudu, A., Son, S., Harjes, E., Spearman, P., Matsuo, H., et al., (2011). Identification of a single amino acid required for APOBEC3 antiretroviral cytidine deaminase activity. *J. Virol.*, **85**, 5691–5695.
92. Murrell, B., Vollbrecht, T., Guatelli, J., Wertheim, J.O., (2016). The evolutionary histories of antiretroviral proteins SERINC3 and SERINC5 do not support an evolutionary arms race in primates. *J. Virol.*, **90**, 8085–8089.
93. Ohno, S., (1970). Evolution by Gene Duplication. Springer-Verlag Heidelberg, Germany.

94. Nakano, Y., Aso, H., Soper, A., Yamada, E., Moriwaki, M., Juarez-Fernandez, G., et al., (2017). A conflict of interest: the evolutionary arms race between mammalian APOBEC3 and lentiviral Vif. *Retrovirology*, **14**, 31.
95. Sawyer, S.L., Emerman, M., Malik, H.S., (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol.*, **2**, E275.
96. Picard, L., Ganivet, Q., Allatif, O., Cimarelli, A., Guéguen, L., Etienne, L., (2020). DGINN, an automated and highly-flexible pipeline for the Detection of Genetic INNovations on protein-coding genes. *bioRxiv*,. 2020.02.25.964155.
97. Hassan, M.A., Butty, V., Jensen, K.D., Saeij, J.P., (2014). The genetic basis for individual differences in mRNA splicing and APOBEC1 editing activity in murine macrophages. *Genome Res.*, **24**, 377–389.
98. Gu, T., Gatti, D.M., Srivastava, A., Snyder, E.M., Raghupathy, N., Simecek, P., et al., (2016). Genetic architectures of quantitative variation in RNA editing pathways. *Genetics*, **202**, 787–798.
99. Shen, F., Kidd, J.M., (2020). Rapid, paralog-sensitive CNV analysis of 2457 human genomes using QuicK-mer2. *Genes*, **11**
100. Tao, L., Jiang, Z., Xu, M., Xu, T., Liu, Y., (2019). Induction of APOBEC3C facilitates the genotoxic stress-mediated cytotoxicity of artesunate. *Chem. Res. Toxicol.*, **32**, 2526–2537.
101. Athanassiou, M., Hu, Y., Jing, L., Houle, B., Zarbl, H., Mikheev, A.M., (1999). Stabilization and reactivation of the p53 tumor suppressor protein in nontumorigenic revertants of HeLa cervical cancer cells. *Cell Growth Diff.: Mol. Biol. J. Am. Assoc. Cancer Res.*, **10**, 729–737.
102. Dull, T., Zufferey, R., Kelly, M., Mandel, R.J., Nguyen, M., Trono, D., et al., (1998). A third-generation lentivirus vector with a conditional packaging system. *J. Virol.*, **72**, 8463–8471.
103. Bähr, A., Singer, A., Hain, A., Vasudevan, A.A., Schilling, M., Reh, J., et al., (2016). Interferon but not MxB inhibits foamy retroviruses. *Virology*, **488**, 51–60.
104. Russell, R.A., Wiegand, H.L., Moore, M.D., Schafer, A., McClure, M.O., Cullen, B.R., (2005). Foamy virus Bet proteins function as novel inhibitors of the APOBEC3 family of innate antiretroviral defense factors. *J. Virol.*, **79**, 8724–8731.
105. Simm, M., Shahabuddin, M., Chao, W., Allan, J.S., Volsky, D.J., (1995). Aberrant Gag protein composition of a human immunodeficiency virus type 1 vif mutant produced in primary lymphocytes. *J. Virol.*, **69**, 4582–4586.
106. Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., et al., (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.*, **40**, 1666–1683.
107. Rose, P.P., Korber, B.T., (2000). Detecting hypermutations in viral sequences with an emphasis on G –> A hypermutation. *Bioinformatics*, **16**, 400–401.
108. Jaguva Vasudevan, A.A., Perkovic, M., Bulliard, Y., Cichutek, K., Trono, D., Häussinger, D., et al., (2013). Prototype foamy virus Bet impairs the dimerization and cytosolic solubility of human APOBEC3G. *J. Virol.*, **87**, 9030–9040.
109. Robert, X., Gouet, P., (2014). Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.*, **42**, W320–W324.
110. Release S. 2: Maestro, Schrödinger, LLC, New York, NY, 2017 (Received: February 2016; 21: 2018.).
111. Shi, K., Carpenter, M.A., Banerjee, S., Shaban, N.M., Kurahashi, K., Salamango, D.J., et al., (2017). Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. *Nature Struct. Mol. Biol.*, **24**, 131–139.
112. Maiti, A., Myint, W., Kanai, T., Delviks-Frankenberry, K., Sierra Rodriguez, C., Pathak, V.K., et al., (2018). Crystal structure of the catalytic domain of HIV-1 restriction factor APOBEC3G in complex with ssDNA. *Nature Commun.*, **9**, 2460.
113. Bas, D.C., Rogers, D.M., Jensen, J.H., (2008). Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins.*, **73**, 765–783.
114. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L., (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
115. D.A. Case VB, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, P.A. Kollman, AMBER 14. University of California, San Francisco, 2014.
116. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., Simmerling, C., (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, **11**, 3696–3713.
117. Li, P., Roberts, B.P., Chakravorty, D.K., Merz Jr., K.M., (2013). Rational design of particle Mesh Ewald compatible Lennard-Jones parameters for +2 metal cations in explicit solvent. *J. Chem. Theory Comput.*, **9**, 2733–2748.
118. Darden, T., York, D., Pedersen, L., (1993). Particle mesh Ewald: An Nṡ log (N) method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.
119. Ryckaert, J.-P., Ciccotti, G., Berendsen, H.J., (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
120. Hopkins, C.W., Le Grand, S., Walker, R.C., Roitberg, A. E., (2015). Long-time-step molecular dynamics through hydrogen mass repartitioning. *J. Chem. Theory Comput.*, **11**, 1864–1874.
121. Zgarbova, M., Sponer, J., Otyepka, M., Cheatham 3rd, T. E., Galindo-Murillo, R., Jurecka, P., (2015). Refinement of the sugar-phosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA. *J. Chem. Theory Comput.*, **11**, 5723–5736.
122. Roe, D.R., Cheatham 3rd., T.E., (2013). PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.*, **9**, 3084–3095.
123. Iwatani, Y., Takeuchi, H., Strebel, K., Levin, J.G., (2006). Biochemical activities of highly purified, catalytically active human APOBEC3G: correlation with antiviral effect. *J. Virol.*, **80**, 5992–6002.
124. Katoh, K., Standley, D.M., (2013). MAFFT multiple sequence alignment software version 7: improvements

in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

125. Stamatakis, A., (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

126. Huson, D.H., Scornavacca, C., (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *System. Biol.*, **61**, 1061–1067.

Bibliography

Bibliography

Agashe, D., Martinez-Gomez, N. C., Drummond, D. A., & Marx, C. J. (2013). Good codons, bad transcript: Large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Molecular Biology and Evolution*, *30*(3), 549–560. https://doi.org/10.1093/molbev/mss273

Agashe, D., Sane, M., Phalnikar, K., Diwan, G. D., Habibullah, A., Martinez-Gomez, N. C., … Marx, C. J. (2016). Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium. *Molecular Biology and Evolution*, *33*(6), 1542–1553. https://doi.org/10.1093/molbev/msw035

Agris, P. F. (1991). Wobble position modified nucleosides evolved to select transfer RNA codon recognition: A modified-wobble hypothesis. *Biochimie*, *73*(11), 1345–1349. https://doi.org/10.1016/0300-9084(91)90163-U

Agris, Paul F., Eruysal, E. R., Narendran, A., Väre, V. Y. P., Vangaveti, S., & Ranganathan, S. V. (2018). Celebrating wobble decoding: Half a century and still much is new. *RNA Biology*, *15*(4–5), 537–553. https://doi.org/10.1080/15476286.2017.1356562

Allan Drummond, D., & Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics*, *10*(10), 715–724. https://doi.org/10.1038/nrg2662

Alonso, A. M., & Diambra, L. (2020). SARS-CoV-2 Codon Usage Bias Downregulates Host Expressed Genes With Similar Codon Usage. *Frontiers in Cell and Developmental Biology*, *8*(August), 1–8. https://doi.org/10.3389/fcell.2020.00831

Amorós-Moya, D., Bedhomme, S., Hermann, M., & Bravo, I. G. (2010). Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage. *Molecular Biology and Evolution*, *27*(9), 2141–2151. https://doi.org/10.1093/molbev/msq103

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), 4310–4315. https://doi.org/10.1186/gb-2010-11-10-r106

Anderson, S., Bankier, A. T., Barrell, B. G., De Bruijn, M. H. L., Coulson, A. R., Drouin, J., … Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, *290*(5806), 457–465. https://doi.org/10.1038/290457a0

Angulo, M., & Carvajal-Rodríguez, A. (2007). Evidence of recombination within human alpha-papillomavirus. *Virology Journal*, *4*, 1–8. https://doi.org/10.1186/1743-422X-4-33

Ankney, J. A., Muneer, A., & Chen, X. (2018). Relative and Absolute Quantitation in Mass Spectrometry-Based Proteomics. *Annual Review of Analytical Chemistry*, *11*, 49–77. https://doi.org/10.1146/annurev-anchem-061516-045357

Atkins, J. F., Wills, N. M., Loughran, G., Wu, C. Y., Parsawar, K., Ryan, M. D., … Nelson, C. C. (2007). A case for “StopGo”: Reprogramming translation to augment codon meaning of GGN by promoting unconventional termination (Stop) after addition of glycine and then allowing continued translation (Go). *Rna*, *13*(6), 803–810. https://doi.org/10.1261/rna.487907

B. Miller, J., Hippen, A. A., M. Wright, S., Morris, C., & G. Ridge, P. (2017). Human viruses have codon usage biases that match highly expressed proteins in the tissues they infect. *Biomedical Genetics and Genomics*, *2*(2), 1–5. https://doi.org/10.15761/bgg.1000134

Bahir, I., Fromer, M., Prat, Y., & Linial, M. (2009). Viral adaptation to host: A proteome-based analysis of codon usage and amino acid preferences. *Molecular Systems Biology*, *5*(311), 1–14. https://doi.org/10.1038/msb.2009.71

Bailey, S. F., Hinz, A., & Kassen, R. (2014). Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. *Nature Communications*, *5*(May), 1–7. https://doi.org/10.1038/ncomms5076

Beck, E., Ludwig, G., Auerswald, E. A., Reiss, B., & Schaller, H. (1982). Nucleotide sequence and exact localization of the neomycin phosphotransferase gene from transposon Tn5. *Gene*, *19*(3), 327–336. https://doi.org/10.1016/0378-1119(82)90023-3

Belalov, I. S., & Lukashev, A. N. (2013). Causes and Implications of Codon Usage Bias in RNA Viruses. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0056642

Bengis, R. G., Leighton, F. A., Fischer, J. R., Artois, M., Mörner, T., & Tate, C. M. (2004). The role of wildlife in emerging and re-emerging zoonoses. *OIE Revue Scientifique et Technique*, *23*(2), 497–511. https://doi.org/10.20506/rst.23.2.1498

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics*, *15*(3), 163–175. https://doi.org/10.1038/nrg3662

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(8), 3171–3175. https://doi.org/10.1073/pnas.74.8.3171

Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., … Münch, C. (2020). Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature*, *583*(7816), 469–472. https://doi.org/10.1038/s41586-020-2332-7

Bourret, J., Alizon, S., & Bravo, I. G. (2019). COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. *Genome Biology and Evolution*, *11*(12), 3523–3528. https://doi.org/10.1093/gbe/evz262

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, *72*(1–2), 248–254. https://doi.org/10.1016/0003-2697(76)90527-3

Bravo, I. G., & Alonso, Á. (2004). Mucosal Human Papillomaviruses Encode Four Different E5 Proteins Whose Chemistry and Phylogeny Correlate with Malignant or Benign Growth. *Journal of Virology*, *78*(24), 13613–13626. https://doi.org/10.1128/jvi.78.24.13613-13626.2004

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Brenner, S., Stretton, A. O. W., & Kaplan, S. (1965). Genetic code: The "nonsense" triplets for chain termination and their suppression. *Nature*, *206*(4988), 994–998. https://doi.org/10.1038/206994a0

Brown, J. A., Valenstein, M. L., Yario, T. A., Tycowski, K. T., & Steitz, J. A. (2012). Formation of triple-helical structures by the 3′-end sequences of MALAT1 and MENβ noncoding RNAs.

Proceedings of the National Academy of Sciences of the United States of America, *109*(47), 19202–19207. https://doi.org/10.1073/pnas.1217338109

Buchan, J. R., & Stansfield, I. (2007). Halting a cellular production line: responses to ribosomal pausing during translation. *Biology of the Cell*, *99*(9), 475–487. https://doi.org/10.1042/bc20070037

Carbone, A., Zinovyev, A., & Képès, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, *19*(16), 2005–2015. https://doi.org/10.1093/bioinformatics/btg272

Chargaff, E., Vischer, E., Doniger, R., Green, C., & Misani, F. (1949). The composition of the desoxypentose nucleic acids of thymus and spleen. *The Journal of Biological Chemistry*, *177*(1), 405–416. https://doi.org/10.1016/s0021-9258(18)57098-8

Chow, L. T., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (1977). An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell*, *12*(1), 1–8. https://doi.org/10.1016/0092-8674(77)90180-5

Chu, D., & Von Der Haar, T. (2012). The architecture of eukaryotic translation. *Nucleic Acids Research*, *40*(20), 10098–10106. https://doi.org/10.1093/nar/gks825

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. *Statistical Power Analysis for the Behavioral Sciences* (second). New York: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Cormack, B. P., Valdivia, R. H., & Falkow, S. (1996). FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, *173*(1), 33–38. https://doi.org/10.1016/0378-1119(95)00685-0

Crick, F. H. C. (1966). Codon—anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology*, *19*(2), 548–555. https://doi.org/10.1016/S0022-2836(66)80022-0

Crick, F. H. C., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, *192*(4809), 1227–1232. https://doi.org/10.1038/1921227a0

Crick, F., & Watson, J. (1953). Molecular Strucutre of Nucleic Acids. *Nature*, *171*, 737–738. Retrieved from https://www.nature.com/articles/171737a0

Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics*, *122*(6), 565–581. https://doi.org/10.1007/s00439-007-0433-0

Daron, J., & Bravo, I. G. (2021). Variability in Codon Usage in Coronaviruses Is Mainly Driven by Mutational Bias and Selective Constraints on CpG Dinucleotide. *Viruses*, *13*(9), 1800. https://doi.org/10.3390/v13091800

Deatherage, D. E., & Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods in Molecular Biology*, *1151*(Ldi), 165–188. https://doi.org/10.1007/978-1-4939-0554-6_12

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*, *25*(10), 1010–1022. https://doi.org/10.1101/gad.2037511

Dekel, E., & Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. *Nature*, *436*(7050), 588–592. https://doi.org/10.1038/nature03842

Desmet, F. O., Hamroun, D., Lalande, M., Collod-Bëroud, G., Claustres, M., & Béroud, C. (2009). Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*, *37*(9), 1–14. https://doi.org/10.1093/nar/gkp215

Dilucca, M., Cimini, G., & Giansanti, A. (2018). Essentiality, conservation, evolutionary pressure and codon bias in bacterial genomes. *Gene*, *663*(October 2017), 178–188. https://doi.org/10.1016/j.gene.2018.04.017

Duret, L. (2002). Evolution of synonymous codon usage in metazoans Duret 641, 640–649.

Eaton, J. D., & West, S. (2020). Termination of Transcription by RNA Polymerase II: BOOM! *Trends in Genetics*, *36*(9), 664–675. https://doi.org/10.1016/j.tig.2020.05.008

Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. *Proceedings of the Royal Society B: Biological Sciences*, *252*(1335), 237–243. https://doi.org/10.1098/rspb.1993.0071

Félez-Sánchez, M., Semeier, J. H. T., Bedhomme, S., González-Bravo, M. I., Kamp, C., & Bravo, I. G. (2015). Cancer, warts, or asymptomatic infections: Clinical presentation matches codon usage preferences in human papillomaviruses. *Genome Biology and Evolution*, *7*(8), 2117–2135. https://doi.org/10.1093/gbe/evv129

Firczuk, H., Kannambath, S., Pahle, J., Claydon, A., Beynon, R., Duncan, J., … McCarthy, J. E. (2013). An in vivo control map for the eukaryotic mRNA translation machinery. *Molecular Systems Biology*, *9*(635), 1–13. https://doi.org/10.1038/msb.2012.73

Forman, D., De Martel, C., Lacey, C. J., Soerjomataram, I., Lortet-Tieulent, J., Bruni, L., … Franceschi, S. (2012). Global Burden of Human Papillomavirus and Related Diseases. *Vaccine*, *30*, 12–23. https://doi.org/10.1016/j.vaccine.2012.07.055

Fricker, L. D. (2015). Limitations of Mass Spectrometry-Based Peptidomic Approaches. *Journal of the American Society for Mass Spectrometry*, *26*(12), 1981–1991. https://doi.org/10.1007/s13361-015-1231-x

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., & Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. *PNAS*, 1–10. https://doi.org/10.1073/pnas.1719375115

Frye, M., T. Haranda, B., Behm, M., & He, C. (2018). RNA modifications modulate gene expression during development. *Science*, *361*(September), 1346–1349.

Ganini, D., Leinisch, F., Kumar, A., Jiang, J. J., Tokar, E. J., Malone, C. C., … Mason, R. P. (2017). Fluorescent proteins such as eGFP lead to catalytic oxidative stress in cells. *Redox Biology*, *12*(March), 462–468. https://doi.org/10.1016/j.redox.2017.03.002

Gardner, M. L., & Freitas, M. A. (2020). Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-up Proteomics, *08*(01), 190–196. https://doi.org/https://doi.org/10.1101/2020.06.29.178335

Gatignol, A., Durand, H., & Tiraby, G. (1988). Bleomycin resistance conferred by a drug-binding protein. *FEBS Letters*, *230*(1–2), 171–175. https://doi.org/10.1016/0014-5793(88)80665-3

Gerrish, P., & Lenski, R. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica*, *102*(0), 127–144. https://doi.org/10.1023/A:1017067816551

Getzenberg, R. H. (1994). Nuclear matrix and the regulation of gene expression: Tissue specificity. *Journal of Cellular Biochemistry*, *55*(1), 22–31. https://doi.org/10.1002/jcb.240550105

Gottschling, M., Bravo, I. G., Schulz, E., Bracho, M. A., Deaville, R., Jepson, P. D., … Nindl, I. (2011). Modular organizations of novel cetacean papillomaviruses. *Molecular Phylogenetics and Evolution*, *59*(1), 34–42. https://doi.org/10.1016/j.ympev.2010.12.013

Graham, F. L., Smiley, J., Russell, W. C., & Nairn, R. (1977). Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *Journal of General Virology*, *36*(1), 59–72. https://doi.org/10.1099/0022-1317-36-1-59

Graille, M., & Séraphin, B. (2012). Surveillance pathways rescuing eukaryotic ribosomes lost in translation. *Nature Reviews Molecular Cell Biology*, *13*(11), 727–735. https://doi.org/10.1038/nrm3457

Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pavé, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, *8*(1), 197. https://doi.org/10.1093/nar/8.1.197-c

Gruber, A. R., Bernhart, S. H., & Lorenz, R. (2015). The viennaRNA web services. *Methods in Molecular Biology*, *1269*, 307–326. https://doi.org/10.1007/978-1-4939-2291-8_19

Harigaya, Y., & Parker, R. (2010). No-go decay: A quality control mechanism for RNA in translation. *Wiley Interdisciplinary Reviews: RNA*, *1*(1), 132–141. https://doi.org/10.1002/wrna.17

Hawkin, J. D. (1988). A survey on intron and exon lengths. *Nucleic Acids Research*, *16*(21), 9893–9908. https://doi.org/10.1093/nar/16.21.9893

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, *2*(1), 13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335

Jeacock, L., Faria, J., & Horn, D. (2018). Codon usage bias controls mRNA and protein abundance in trypanosomatids. *ELife*, *7*. https://doi.org/10.7554/eLife.32496

Jiao, X., Xiang, S., Oh, C., Martin, C. E., Tong, L., & Kiledjian, M. (2010). Identification of a quality-control mechanism for mRNA 5′-end capping. *Nature*, *467*(7315), 608–611. https://doi.org/10.1038/nature09338

Johnson, C. K., Hitchens, P. L., Pandit, P. S., Rushmore, J., Evans, T. S., Young, C. C. W., & Doyle, M. M. (2020). Global shifts in mammalian population trends reveal key predictors of virus spillover risk. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1924). https://doi.org/10.1098/rspb.2019.2736

Jones, B. A., Grace, D., Kock, R., Alonso, S., Rushton, J., Said, M. Y., … Pfeiffer, D. U. (2013). Zoonosis emergence linked to agricultural intensification and environmental change. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8399–8404. https://doi.org/10.1073/pnas.1208059110

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Research*, *12*(6), 962–968. https://doi.org/10.1101/gr.87702

Jukes, T. H., & Osawa, S. (1993). Evolutionary changes in the genetic code. *Comparative Biochemistry and Physiology -- Part B: Biochemistry And*, *106*(3), 489–494. https://doi.org/10.1016/0305-0491(93)90122-L

Juven-Gershon, T., Hsu, J. Y., Theisen, J. W., & Kadonaga, J. T. (2008). The RNA polymerase II core promoter - the gateway to transcription. *Current Opinion in Cell Biology*, *20*(3), 253–259. https://doi.org/10.1016/j.ceb.2008.03.003

Kane, J. F. (1995). Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Current Opinion in Biotechnology*, *6*(5), 494–500. https://doi.org/10.1016/0958-1669(95)80082-4

Kaufmann, W. K., & Paules, R. S. (1996). DNA damage and cell cycle checkpoints. *The FASEB Journal*, *10*(2), 238–247. https://doi.org/10.1096/fasebj.10.2.8641557

Kaur, J., Kumar, A., & Kaur, J. (2018). Strategies for optimization of heterologous protein expression in E. coli: Roadblocks and reinforcements. *International Journal of Biological Macromolecules*, *106*, 803–822. https://doi.org/10.1016/j.ijbiomac.2017.08.080

Khorana, H. G., Büchi, H., Ghosh, H., Gupta, N., Jacob, T. M., Kössel, H., … Wells, R. D. (1966). Polynucleotide synthesis and the genetic code. *Cold Spring Harbor Symposia on Quantitative Biology*, *31*, 39–49. https://doi.org/10.1101/SQB.1966.031.01.010

Kim, S. J., Yoon, J. S., Shishido, H., Yang, Z., Rooney, L. A. A., Barral, J. M., & Skach, W. R. (2015). Translational tuning optimizes nascent protein folding in cells. *Science*, *348*(6233), 444–448. https://doi.org/10.1126/science.aaa3974

Lachapelle, J., Reid, J., & Colegrave, N. (2015). Repeatability of adaptation in experimental populations of different sizes. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1805). https://doi.org/10.1098/rspb.2014.3033

Lara-Ramírez, E. E., Salazar, M. I., López-López, M. D. J., Salas-Benito, J. S., Sánchez-Varela, A., & Guo, X. (2014). Large-scale genomic analysis of codon usage in dengue virus and evaluation of its phylogenetic dependence. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/851425

Lebeuf-Taylor, E., McCloskey, N., Bailey, S. F., Hinz, A., & Kassen, R. (2019). The distribution of fitness effects among synonymous mutations in a gene under directional selection. *ELife*, *8*, 1–16. https://doi.org/10.7554/eLife.45952

Li, Y., & Kiledjian, M. (2010). Regulation of mRNA decapping. *Wiley Interdisciplinary Reviews: RNA*, *1*(2), 253–265. https://doi.org/10.1002/wrna.15

Lin, D., Li, L., Xie, T., Yin, Q., Saksena, N., Wu, R., … Chen, X. (2018). Codon usage variation of Zika virus: The potential roles of NS2B and NS4A in its global pandemic. *Virus Research*, *247*, 71–83. https://doi.org/10.1016/j.virusres.2018.01.014

Liu, Y. (2020). A code within the genetic code: Codon usage regulates co-translational protein folding. *Cell Communication and Signaling*, *18*(1), 1–9. https://doi.org/10.1186/s12964-020-00642-6

Lujan, S. A., Williams, J. S., Pursell, Z. F., Abdulovic-Cui, A. A., Clark, A. B., Nick McElhinny, S. A., & Kunkel, T. A. (2012). Mismatch Repair Balances Leading and Lagging Strand DNA Replication Fidelity. *PLoS Genetics*, *8*(10). https://doi.org/10.1371/journal.pgen.1003016

McGary, K., & Nudler, E. (2013). RNA polymerase and the ribosome: The close relationship. *Current Opinion in Microbiology*, *16*(2), 112–117. https://doi.org/10.1016/j.mib.2013.01.010

Merrick, W. C. (2004). Cap-dependent and cap-independent translation in eukaryotic systems. *Gene*, *332*(1–2), 1–11. https://doi.org/10.1016/j.gene.2004.02.051

Meyer, S., Temme, C., & Wahle, E. (2004). Messenger RNA turnover in eukaryotes: Pathways and enzymes. *Critical Reviews in Biochemistry and Molecular Biology*, *39*(4), 197–216. https://doi.org/10.1080/10409230490513991

Mitton-Fry, R. M., DeGregorio, S. J., Wang, J., Steitz, T. A., & Steitz, J. A. (2010). Poly(A) tail recognition by a viral RNA element through assembly of a triple helix. *Science*, *330*(6008), 1244–1247. https://doi.org/10.1126/science.1195858

Mordstein, C., Cano, L., Morales, A. C., Young, B., Ho, A. T., Rice, A. M., … Kudla, G. (2021). Transcription, mRNA Export, and Immune Evasion Shape the Codon Usage of Viruses. *Genome Biology and Evolution*, *13*(9), 1–14. https://doi.org/10.1093/gbe/evab106

Müller, M. M., Gerster, T., & Schaffner, W. (1988). Enhancer sequences and the regulation of gene transcription. *European Journal of Biochemistry*, *176*(3), 485–495. https://doi.org/10.1111/j.1432-1033.1988.tb14306.x

Narechania, A., Chen, Z., DeSalle, R., & Burk, R. D. (2005). Phylogenetic Incongruence among Oncogenic Genital Alpha Human Papillomaviruses. *Journal of Virology*, *79*(24), 15503–15510. https://doi.org/10.1128/jvi.79.24.15503-15510.2005

Németh, A., Perez-Fernandez, J., Merkl, P., Hamperl, S., Gerber, J., Griesenbeck, J., & Tschochner, H. (2013). RNA polymerase I termination: Where is the end? *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1829*(3–4), 306–317. https://doi.org/10.1016/j.bbagrm.2012.10.007

Ngoka, L. C. M. (2008). Sample prep for proteomics of breast cancer: Proteomics and gene ontology reveal dramatic differences in protein solubilization preferences of radioimmunoprecipitation assay and urea lysis buffers. *Proteome Science*, *6*, 1–24. https://doi.org/10.1186/1477-5956-6-30

Nirenberg, M. W., & J. Heinrich Matthaei. (1961). The dependence of cell- free protein synthesis in E. coli upon naturally occuring or synthetic polyribonucleotides, 1588–1602.

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. *Annual Review of Biochemistry*, *79*(1), 321–349. https://doi.org/10.1146/annurev-biochem-060208-105251

Ohno, K., Takeda, J. I., & Masuda, A. (2018). Rules and tools to predict the splicing effects of exonic and intronic mutations. *Wiley Interdisciplinary Reviews: RNA*, *9*(1). https://doi.org/10.1002/wrna.1451

Pasleau, F., Tocci, M. J., Leung, F., & Kopchick, J. J. (1985). Growth hormone gene expression in eukaryotic cells directed by the Rous sarcoma virus long terminal repeat or cytomegalovirus immediate-early promoter. *Gene*, *38*(1–3), 227–232. https://doi.org/10.1016/0378-1119(85)90221-5

Payne, B. L., & Alvarez-Ponce, D. (2019). Codon usage differences among genes expressed in different tissues of drosophila melanogaster. *Genome Biology and Evolution*, *11*(4), 1054–1065. https://doi.org/10.1093/gbe/evz051

Perona, J. J., & Hadd, A. (2012). Structural diversity and protein engineering of the aminoacyl-tRNA Synthetases. *Biochemistry*, *51*(44), 8705–8729. https://doi.org/10.1021/bi301180x

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *National Review of Genetics*, *12*(1), 32–42. https://doi.org/10.1038/nrg2899.Synonymous

Plotkin, J. B., Robins, H., & Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(34), 12588–12591. https://doi.org/10.1073/pnas.0404957101

Pouyet, F., Mouchiroud, D., Duret, L., & Sémon, M. (2017). Recombination, meiotic expression and human codon usage. *ELife*, *6*, 1–19. https://doi.org/10.7554/eLife.27344

Poverennaya, I. V., & Roytberg, M. A. (2020). Spliceosomal Introns: Features, Functions, and Evolution. *Biochemistry (Moscow)*, *85*(7), 725–734. https://doi.org/10.1134/S0006297920070019

Puigbò, P., Guzmán, E., Romeu, A., & Garcia-Vallvé, S. (2007). OPTIMIZER: A web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Research*, *35*(SUPPL.2), 126–131. https://doi.org/10.1093/nar/gkm219

Quax, T. E. F., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, *59*(2), 149–161. https://doi.org/10.1016/j.molcel.2015.05.035

Racle, J., Picard, F., Girbal, L., Cocaign-Bousquet, M., & Hatzimanikatis, V. (2013). A Genome-Scale Integration and Analysis of Lactococcus lactis Translation Data. *PLoS Computational Biology*, *9*(10). https://doi.org/10.1371/journal.pcbi.1003240

Rector, A., Stevens, H., Lacave, G., Lemey, P., Mostmans, S., Salbany, A., … Van Ranst, M. (2008). Genomic characterization of novel dolphin papillomaviruses provides indications for recombination within the Papillomaviridae. *Virology*, *378*(1), 151–161. https://doi.org/10.1016/j.virol.2008.05.020

Reenan, R. A. (2001). The RNA world meets behavior: Adenosine-to-inosine pre-mRNA editing in animals. *Trends in Genetics*, *17*(2), 53–56. https://doi.org/10.1016/S0168-9525(00)02169-7

Robinson, F., Jackson, R. J., & Smith, C. W. J. (2008). Expression of human nPTB is limited by extreme suboptimal codon content. *PLoS ONE*, *3*(3). https://doi.org/10.1371/journal.pone.0001801

Robles-Sikisaka, R., Rivera, R., Nollens, H. H., St. Leger, J., Durden, W. N., Stolen, M., … Wellehan, J. F. X. (2012). Evidence of recombination and positive selection in cetacean papillomaviruses. *Virology*, *427*(2), 189–197. https://doi.org/10.1016/j.virol.2012.01.039

Rohr, J. R., Barrett, C. B., Civitello, D. J., Craft, M. E., Delius, B., DeLeo, G. A., … Tilman, D. (2019). Emerging human infectious diseases and the links to global food production. *Nature Sustainability*, *2*(6), 445–456. https://doi.org/10.1038/s41893-019-0293-3

Roundtree, I. A., Evans, M. E., Pan, T., & He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation. *Cell*, *169*(7), 1187–1200. https://doi.org/10.1016/j.cell.2017.05.045

Ryabova, L. A., Pooggin, M. M., & Hohn, T. (2006). Translation reinitiation and leaky scanning in plant viruses. *Virus Research*, *119*(1), 52–62. https://doi.org/10.1016/j.virusres.2005.10.017

Saikia, M., Wang, X., Mao, Y., Wan, J., Pan, T., & Qian, S. B. (2016). Codon optimality controls differential mRNA translation during amino acid starvation. *RNA*, *22*(11), 1719–1727. https://doi.org/10.1261/rna.058180.116

Sawilowsky, S. S. (2009). Very large and huge effect sizes. *Journal of Modern Applied Statistical Methods*, *8*(2), 597–599. https://doi.org/10.22237/jmasm/1257035100

Schoustra, S. E., Bataillon, T., Gifford, D. R., & Kassen, R. (2009). The properties of adaptive walks in evolving populations of fungus. *PLoS Biology*, *7*(11). https://doi.org/10.1371/journal.pbio.1000250

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., & Plotkin, J. B. (2013). XRate-limiting steps in yeast protein translation. *Cell*, *153*(7), 1589. https://doi.org/10.1016/j.cell.2013.05.049

Sharp, P. M., & Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295. https://doi.org/10.1093/nar/15.3.1281

Sikic, B. I. (1986). Biochemical and cellular determinants of bleomycin cytotoxicity. *Cancer Surveys*, *5*(1), 81–91.

Simmonds, P. (2020). Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *MSphere*, *5*(3), 1–13. https://doi.org/10.1128/msphere.00408-20

Smale, S. T., & Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual Review of Biochemistry*, *72*, 449–479. https://doi.org/10.1146/annurev.biochem.72.121801.161520

Smith, A. M., Abu-Shumays, R., Akeson, M., & Bernick, D. L. (2015). Capture, unfolding, and detection of individual tRNA molecules using a nanopore device. *Frontiers in Bioengineering and Biotechnology*, *3*(JUN), 1–11. https://doi.org/10.3389/fbioe.2015.00091

Szendro, I. G., Franke, J., De Visser, J. A. G. M., & Krug, J. (2013). Predictability of evolution depends nonmonotonically on population size. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(2), 571–576. https://doi.org/10.1073/pnas.1213613110

Tawfik, D. S., & Gruic-Sovulj, I. (2020). How evolution shapes enzyme selectivity – lessons from aminoacyl-tRNA synthetases and other amino acid utilizing enzymes. *FEBS Journal*, *287*(7), 1284–1305. https://doi.org/10.1111/febs.15199

Thomas, P., & Smart, T. G. (2005). HEK293 cell line: A vehicle for the expression of recombinant proteins. *Journal of Pharmacological and Toxicological Methods*, *51*(3 SPEC. ISS.), 187–200. https://doi.org/10.1016/j.vascn.2004.08.014

Torabi, S. F., Vaidya, A. T., Tycowski, K. T., DeGregorio, S. J., Wang, J., Shu, M. Di, … Steitz, J. A. (2021). RNA stabilization by a poly(A) tail 3'-end binding pocket and other modes of poly(A)-RNA interaction. *Science*, *371*(6529). https://doi.org/10.1126/science.abe6523

Torrent, M., Chalancon, G., De Groot, N. S., Wuster, A., & Madan Babu, M. (2018). Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science Signaling*, *11*(546), 1–10. https://doi.org/10.1126/scisignal.aat6409

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., … Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, *141*(2), 344–354. https://doi.org/10.1016/j.cell.2010.03.031

Verosloff, M. S., Corcoran, W. K., Dolberg, T. B., Bushhouse, D. Z., Leonard, J. N., & Lucks, J. B. (2021). RNA Sequence and Structure Determinants of Pol III Transcriptional Termination in Human Cells. *Journal of Molecular Biology*, *433*(13), 166978. https://doi.org/10.1016/j.jmb.2021.166978

Voet, D., Voet, J. G., & Pratt, C. W. (2016). *Fundamentals of Biochemistry: Life at the Molecular Level*. Wiley. Retrieved from https://books.google.fr/books?id=9T7hCgAAQBAJ

Wada, K. nosuke, Aota, S. ichi, Tsuchiya, R., Ishibashi, F., Gojobori, T., & Ikemura, T. (1990). Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Research*, *18*(01656004), 2367–2411. https://doi.org/10.1093/nar/18.suppl.2367

Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A., & Clark, P. L. (2020). Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell

fitness. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(7), 3528–3534. https://doi.org/10.1073/pnas.1907126117

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., & Ni, Y. (2018). Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-017-19120-0

Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., & Bartel, D. P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Reports*, *14*(7), 1787–1799. https://doi.org/10.1016/j.celrep.2016.01.043

Weinreich, D. M., Watson, R. A., & Chao, L. (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, *59*(6), 1165–1174. https://doi.org/10.1111/j.0014-3820.2005.tb01768.x

Wolfe, K. H., Sharp, P. M., & Li, W. H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, *337*(6204), 283–285. https://doi.org/10.1038/337283a0

Wright, F. (1990). The “effective number of codons” used in a gene. *Gene*, *87*(1), 23–29. https://doi.org/10.1016/0378-1119(90)90491-9

Abstract

Eukaryotic cells contain a complex cellular machinery, that regulates and carries out gene expression. The standard genetic code that is the basis of this protein production line is redundant, meaning that 64 codons encode for 20 amino acids. This redundancy gives rise to synonymous codons, that encode for the same amino acid. Synonymous codons are not used at random, genes, tissues and organisms tend to have divergent Codon Usage Preferences (CUPrefs). The role of CUPrefs and the forces that shape them are not yet clear, although it is certain that they hold an important regulatory position in gene expression. If a gene's CUPrefs match the cellular tRNA pool, translation will be fast and efficient, while under- or overmatching CUPrefs may cause either slow and inaccurate translation or competition among genes for resources. Viruses are dependent of the host cell's resources to express their genes, therefore the study of their CUPrefs is primordial to understand their functioning and interactions with the host. In this work, we attempt to enlarge our understanding of the importance of CUPrefs by analyzing the causes and consequences of CUPrefs in eukaryotes and viruses, and in a long-term evolution experiment.

First, we analyzed eleven recombinant Papillomaviruses (PV) that infect exclusively Cetaceans, along with other PVs that infect the same host order: the Cetartiodactyles. We found that recombinant PVs, are not different from non-recombinants in terms of CUPrefs. Instead CUPrefs are associated to gene type, with a link to gene function, and expression pattern. They do not match host CUPrefs, hinting to an immune evasion strategy by keeping low viral gene expression due to the undermatch. Next, we looked at the evolution of CUPrefs in the three paralogs in vertebrates encoding for the Polypyrimidin tract binding protein (PTBP). The PTBP paralogs show distinct CUPrefs, with a GC enrichment linked to local mutational forces in PTBP1 in mammals. We propose that the divergent nucleotide composition in PTBPs is a result of evolution by sub-functionalisation upon gene duplication, and that it's linked to gene expression patterns in different tissues. In an experimental evolution setup we introduced synonymous genes (that only differ in CUPrefs) under strong selection for expression into HEK293 cells, and let them evolve under three conditions for a hundred generations. When the heterologous genes under are directly under selection, cells overcome CUPrefs mismatch, and in spite of the differences, converge to a similar expression pattern. In contrast, when the modified genes are subject of genetic hitchhiking, regulatory mechanisms lead to different expression profiles to limit metabolic cost.

Overall we show that the CUPrefs play a role in regulating gene expression in terms of its differed time or place. Further, we suggest that Eukaryote cells can adjust rapidly by complex regulatory mechanisms to overcome the disadvantages of heterologous CUPrefs if they are needed for survival, or down-regulate them if their expression is costly.