Analyse statistique de la sélection dans des banques minimalistes de protéines

Sébastien Boyer

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Physique/Physique pour les sciences du vivant**

Arrêté ministériel :

Présentée par

Sébastien Boyer

Thèse dirigée par **Bahram Houchmandzadeh**
et codirigée par **Olivier Rivoire & Clément Nizak**

préparée au sein **Laboratoire Interdisciplinaire de Physique**
et de **École doctorale de Physique de Grenoble**

Analyse statistique de la sélection dans des banques minimalistes de protéines.

Thèse soutenue publiquement le **1er Octobre 2015**,
devant le jury composé de :

M. Hervé Isambert
Institut Curie, Rapporteur
M. Phillipe Minard
IBBMC, Président
M. Eric Bertin
LIPhy, Examinateur
M. Andrew Griffiths
ESPCI, Examinateur
M. Bahram Houchmandzadeh
LIPhy, Directeur de thèse
M. Olivier Rivoire
LIPhy, Co-Directeur de thèse
M. Clément Nizak
ESPCI, Co-Directeur de thèse

BOYER SÉBASTIEN

ANALYSE STATISTIQUE DE LA SÉLECTION DANS DES BANQUES MINIMA-LISTES DE PROTÉINES.

Remerciements

Je voudrais remercier Olivier Rivoire et Clément Nizak pour avoir donné l'opportunité au physicien que je suis de faire cette plongée dans le monde de la biologie et des statistiques. Merci de m'avoir fait confiance du début à la fin et à tous moments : des westerns blots ratés en passant par les contaminations de phages, de l'extension du projet au anticorps anti-HIV jusqu'au projet EVT. Merci Olivier de n'avoir pas craqué quand tous les matins pendant 2 mois je me présentais à ton bureau en te posant des questions existentielles sur la robustesse de mon analyse. Je voudrais remercier Barham Houchmandzadeh, de m'avoir accepté comme son premier étudiant thésard ainsi que pour nos nombreuses discussions toujours palpitantes sur à peu prés tous les sujets. En citant les personnes qui m'ont fait confiance et sans qui je n'aurai pas pu participer à ce magnifique projet, il faut incontestablement citer Irina Mihalscescu sans qui le master Physique pour l'exploration du vivant ne serait pas et qui m'a soutenu jusqu'au bout pour cette bourse de thèse.
Je n'ai bien sur pas été tout seul sur les paillasses de la salle de biologie moléculaire. Merci à Anand Soshee Kumar de m'avoir appris le phage display et Dipanwita Biswas pour la création des banques. Mais surtout, merci à Natale Scarramozino, pour son aide inconditionnelle aussi bien sur la paillasse que lorsque j'avais besoin d'encouragements. Merci à Mathieu Hemery pour avoir été aussi pédagogue lorsque les questions théoriques se faisaient pressantes. Enfin, merci à toute l'équipe administrative du LIPhy qui a été d'une efficacité exemplaire pour tous nos déplacements.
Mais cette thèse elle a aussi et surtout été rendu possible grâce à ma mère qui m'a appris à lire et à compter, à mon père qui m'a fait aimé lire et aimé les sciences en m'introduisant à la Science fiction, et à ma soeur qui explorait le monde des sciences comme une éclaireuse pour moi. Cette thèse elle a pu être en bonne voix, notamment grâce à la famille Simonet, qui m'a pris sous son aile quand je suis arrivé en Métropole et qui aujourd'hui et l'un des plus grands acteurs de l'aboutissement de cette thèse. Enfin, merci à Rachel Genthial, sans qui j'aurais terminé cette thèse dans un piteux état.

Table des matières

Introduction

Historique de l'évolution par sélection naturelle

L'évolution par sélection naturelle est un processus introduit par Charles Darwin pour expliquer les convergences ou divergences phénotypiques observées entre les espèces [1]. Ses observations lors de son voyage à bord de la deuxième expédition de Fitz Roy, ainsi que sa synthèse des connaissances en reproduction sélective de céréales ou d'animaux domestiqués, lui permettent de proposer un mécanisme de filiation et d'évolution des espèces :
"As many more individuals of each species are born than can possibly survive ; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being,if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be naturally selected. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form".
On peut synthétiser ce mécanisme comme un processus en trois étapes contenant chacune leur part de stochasticité : création de diversité, sélection, amplification. A l'époque de *On the Origin of species* on n'a pas caractérisé cette stochasticité et on ne connait ni le mécanisme de transmission de l'hérédité ni la manière dont cette création de diversité apparait.

1. Charles Darwin. *On the Origin of species*. Oxford University Press, Oxford, 2008

Il faut attendre la redécouverte des travaux de Gregor Mendel, effectués entre 1853 et 1866 [2], par Hugo de Vries et Carl Correns (1900) pour sortir des théories d'hérédité par mélange et de pangenése, et aboutir à une théorie discrète de l'hérédité basée sur les trois lois de Mendel . La première loi stipule que chaque individu contient une paire de 'particules' codant pour un trait similaire mais différent, dont une des deux est transmise aléatoirement à sa descendance. La seconde, elle, se concentre sur le fait que des particules codant pour des traits différents peuvent être transmises de façon discrète et indépendante les unes des autres. Enfin, la troisième stipule qu'il peut y avoir des relations de dominances entre les différents allèles d'une même particule , de sorte que même lors d'une reproduction sexuée où un individu reçoit une particule de chacun de ses parents, seulement l'une des deux peut s'exprimer. Dès 1902 Walter Sutton [3] [4] et Theodor Boveri [5] hypothètisent que ces particules si importantes à la Théorie de Mendel seraient les chromosomes. En 1944 Oswald Avery, sur la base de travaux de Frederick Griffith datant de 1928,

Fitness : capacité d'un individu à propager ses gènes.

Allèle : une des multiples versions différentes d'un gène.

3. W.S. Sutton. On the morphology of the chromosome group in brachystola magna. *Biol. Bull.*, 4:24–39, 1902

4. W.S.l Sutton. The chromosomes in heredity. *Biol. Bull.*, 4:231–251, 1903

5. T.H. Boveri. Ergebnisse uber die konstitution der chromatischen substanz des zelkerns. 1904

découvre que l'ADN est bien le support de l'hérédité, en montrant de façon contrôlée que l'absorption par une bactérie d'un morceau d'ADN codant pour un certain trait lui permettait d'obtenir ce trait [6]. Frederick Griffith jette alors les bases d'un des mécanismes de génération de diversité au niveau génétique : les recombinaisons. Un autre mécanisme basé sur les erreurs lors de la duplication de l'ADN est envisagé grâce à l'expérience de Meselson et Stahl en 1958 [7] qui explique le processus de réplication de l'ADN.

A la lecture du paragraphe précédent on peut avoir l'idée trompeuse que toutes ces découvertes se sont enchainées d'une manière fluide et naturelle. Rien n'est plus faux, et il est intéressant de regarder de plus près l'enchainement des deux découvertes majeures que sont le principe de sélection naturelle et le principe d'hérédité mendélien, ainsi que l'ensemble des questions que ces principes ont pu soulever à l'époque [8].

Il est tout d'abord intéressant de constater que hérédité mendélienne et sélection naturelle étaient au départ perçues comme antagonistes, notamment parce que l'hérédité mendélienne réfute l'hypothèse d'hérédité par mélange (à cause de la dominance). Ainsi les scientifiques de l'époque ont tout d'abord embrassé l'hypothèse d'hérédité mendélienne comme significative pour la création de la variabilité plutôt que comme mécanisme d'hérédité. Pour autant entre 1900 et 1918 on commence à envisager que sélection naturelle et hérédité mendélienne sont complémentaires, et de nombreux scientifiques tel que Punnet, Weinberg, Norton, Pearl, Jenning (etc..) se lancent dans la compréhension des implications que l'hérédité mendélienne aurait sur la sélection naturelle. Cette étude est dirigée par 3 questions principales qui seront encore d'actualité lorsque la théorie devra être confrontée à l'expérience un peu moins de 20 ans plus tard. Une des premières questions abordées fut celle de la distribution génétique d'une population soumise à de l'hybridisation (deux génotypes AA et aa, où A et a sont deux allèles codant pour le même trait), et notamment si l'équilibre de cette dynamique menait à la perte de l'hybridisation et un retour au type parental [9]. Dans la même veine, entre 1912 et 1916, Jenning et Pearl se demandaient si le croisement consanguin était un choix judicieux pour la conservation d'un trait d'intérêt. Il était donc déjà question d'identifier les paramètres à utiliser pour diriger/optimiser l'évolution et de l'importance potentielle de la dynamique évolutive sur la réponse du système. Enfin Punnett et Norton se demandèrent quelles sont les échelles de temps mises en jeux pour qu'un génotype envahisse une population, ou disparaisse, et cela en fonction de la pression de sélection.

Une des conclusions de cette étude fut que dans la plupart des cas, la sélection dans le contexte de l'hérédité Mendélienne est un processus très efficace, dans la plupart des configurations génétiques. Pour autant la sélection naturelle associée à l'hérédité mendélienne n'a que très peu la capacité d'éliminer un allèle rare, récessif et délétère. Warren fut l'un des premiers à publier un papier qui conclut explicitement que Darwinisme et hérédité Mendélienne sont com-

6. Oswald T.Avery and al. Studies on the chemical nature of the substance inducing transformation of pneumoccocal types. *J. Exp. Med.*, 79:137–158, 1944

7. M. Meselson and al. The replication of dna in escherichia coli. *PNAS*, 44:671–682, 1958

8. William B. Provine. *The origins of Theoritical Population Genectics*. The university of Chicago Press, 1992

Hérédité par mélange : théorie supposant que le phénotype d'un individu soit la moyenne plus ou moins pondérée du phénotype de ses parents.

9. Yule. Mendel s law and their probable relations to intraracial heredity. page 225

patibles [10]. Encore une fois ce papier se focalise sur la distribution génotypique d'une population à deux allèles.

Ainsi au début de 1918, l'hérédité Mendélienne est acceptée comme le mécanisme d'hérédité qui permet à la sélection naturelle d'être efficace même pour de légères différences entre individus. Mais tout cela relevant du domaine des mathématiques, il reste encore à prouver que la théorie s'accorde à l'expérience, et notamment aux mesures phénotypiques.

C'est dans ce contexte scientifique que le trio fondateur de la génétique des populations et de la théorie synthétique de l'évolution entre en scène dès 1918 : Ronald Alymer Fisher, Sewall Wright, J.B.S. Haldane. Des données biométriques sont à leur disposition et il faut alors développer des outils statistiques nouveaux pour analyser ces données. Il est tout à fait remarquable que nombre de ces outils soient maintenant partis intégrantes de l'arsenal statistique utilisé dans de nombreux domaines (Maximum Likelihood, Pearson correlation, Information de Fisher ....) et les modèles théoriques développés à cette époque sont tout à fait avant-gardistes (on pense par exemple à la notion de paysage de potentiel que Wright introduisit avec son paysage d'adaptation). Par exemple, pour pouvoir comparer les théories mathématiques développées avant 1918 et les résultats d'expériences il faut pouvoir comparer les corrélations obtenues expérimentalement dans les mesures biométriques entre parents issus de l'expérience, et les corrélations hypothétiques qu'impliqueraient sélection et hérédité mendélienne. Fisher s'attela à l'analyse des corrélations phénotypiques entre parents [11]. C'est d'ailleurs au cours de cette étude qu'il introduit la notion de variance comme étant le paramètre important, car c'est la variance (cumulant), et non la déviation standard, qui s'ajoute lorsque l'on s'intéresse à la statistique de deux variables aléatoires. Il pointe notamment du doigt le fait que la variabilité entre parents est plus grande que ce que la simple hérédité pouvait prétendre. Il montre par exemple qu'entre deux frères, 54 pour-cents de leur variance biométrique est due à leur parenté et que le reste est du à d'autres types de variations. Ces autres types de variations étant par exemple : les effets de l'environnement, les interactions entre gênes ou encore la dominance entre allèles. Fisher montre qu'il est possible de faire la différence entre ces différentes variances. Wright aussi aborde la question des multiples causes de variation entre parents en introduisant la méthode des chemins de coefficients . L'utilisation de cette méthode dans l'analyse des effets d'un fort croisement consanguin sur la couleur du pelage des cochons d'inde lui permirent de voir les effets importants de la variation développementale. Encore une fois il est intéressant de noter que déjà les causes de variations non génétiques sont abordées.

En 1922, Fisher publie un condensé de ses recherches [12], où il discute en détail des interactions entre sélection, dominance, mutation, extinction aléatoire de gènes et appariement assortatif. Fischer s'intéresse ainsi aux chances de survie d'une mutation dans une population en fonction de la taille de cette population. La question posée est

10. Howard C. Warren. Numerical effect of natural selection acting upon mendelian characters. *Genetics2*, pages 305–312, 1917

11. R.A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Royal Society of Edinburg*, 1918

12. R.A.Fisher. On the dominance ratio. *Proceeding of the Royal Society of Edinburgh*, 42:321–341, 1922

donc de quantifier l'efficacité de la sélection par rapport à l'aléatoire inhérent à des processus de vie et de mort dans des populations (drift génétique). Sur ce point Fisher et Wright s'opposent, principalement parce qu'ils ne parlent pas du même système ou tout du moins du même régime de la dynamique évolutive : Fisher regarde un allèle en particulier alors que Wright parle d'un gène en interaction avec d'autres gènes. Wright, au contraire de Fisher, pense que le drift génétique est un élément crucial de la dynamique d'évolution, non pas comme antagoniste à la sélection mais plutôt offrant plus de possibilités à la sélection. En effet, Sewall Wright est particulièrement connu pour avoir introduit le concept de paysage d'adaptation (fitness landscape), qui est une représentation dans l'espace des génotypes des effets de fitness . Dans ce contexte, le drift génétique permet de traverser une vallée de faible fitness pour atteindre un pic. La discorde entre les deux scientifiques concernant la place du drift génétique dans l'évolution, se rapportera un peu plus tard, notamment sous l'impulsion de Kimura, à s'interroger sur la taille des effets de mutation.

Avant de plonger dans cette dernière interrogation, qui concerne au premier plan le travail de cette thèse, retournons un moment sur le travail de Fisher et sur la question qui au départ fut à la racine de toutes les autres : l'adaptation est-elle continue ou apparaît elle par saut (pour une revue complète des thèmes abordés ci-après [13]) ? Fisher est partisan de l'hypothèse continue et a élaboré un model dit infinitésimal : les mutations ont de petits effets phénotypiques mais touchent un grand nombre de gènes dont l'effet cumulé peut aboutir à de grands effets de fitness. Dans **The genetical theory of natural selection**, Fisher expose son modèle dit Modèle géométrique de Fisher Fig.1,qu'il utilise pour explorer la question suivante : quelle est la probabilité qu'une mutation entrainant un effet phénotypique de taille s soit bénéfique ? Il montre ainsi que dans les hypothèses de son modèle les grands effets phénotypiques sont très largement peu probables par rapport aux petits effets phénotypiques. Il en conclut donc que l'adaptation procède par petites étapes même si de rares événements de tailles intermédiaires ou grandes peuvent se produire. Des années plus tard un argument de Motoo Kimura [14] viendra mettre à mal la conclusion de Fisher. En effet Kimura fait remarquer que pour être importante dans la dynamique adaptative, il ne suffit pas à une mutation d'être bénéfique mais il lui faut aussi avoir un effet suffisamment grand pour permettre au mutant d'avoir une chance de survivre aux processus de vie et de mort. Selon lui donc, les mutations importantes sont celles de moyenne amplitude au niveau de leurs effets phénotypiques. Un autre reproche fait au modèle géométrique de Fisher est qu'il se place dans l'espace des phénotypes, et non dans la base légitime quand on parle d'hérédité et de mutations, qui est génétique ou tout du moins relative à la séquence d'acide aminé codant pour le phénotype étudié. Avec cette idée Maynard Smith [15] pose les bases de l'évolution moléculaire dans un modèle discret par nature. Reprenant le fitness landscape de

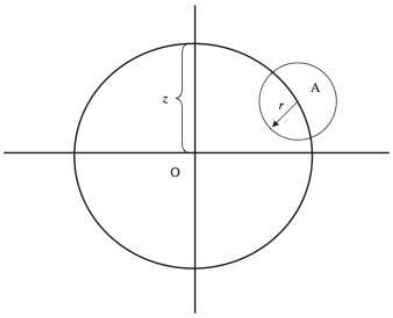


FIGURE 1: **Fisher's geometric model.(Orr, 2005)** Dans cet exemple il n'y a que deux traits représentés par l'axe x et y. L'organisme qui s'adapte (point A) se trouve à une distance z du point o qui est le maximum local d'adaptation. L'organisme peut se déplacer par mutation dans l'espace phénotypique en suivant le vecteur **r** dont la magnitude est fixé mais dont la direction est aléatoire (petit cercle). Toute mutation qui permet d'entrer dans le grand cercle est favorable.

13. H. Allen Orr. The genetic theory of adaptation : a brief history. *Nature review Genetics*, 6:119–127, 2005

14. Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge Univ. Press, 1983

15. Maynard Smith. *The Scientist Speculates: an Anthology of Partly-Baked Ideas*. Basic Books,New York, 1962

Wright dans l'espace des séquences de Maynard Smith, Kauffman et Levin [16], adressent, dans leur NK model Fig.2, les questions suivantes : combien y a-t-il de maximums locaux de fitness, combien d'entre eux peuvent être effectivement atteint et combien de mutations sont nécessaires pour y arriver ? Selon Gillespie, il manque un ingrédient crucial à cette théorie : l'évolution agit sur un organisme ou une protéine déjà plus ou moins adaptée à l'environnement, et donc sonde une zone particulière du fitness landscape. En ajoutant à cela l'hypothèse de Kimura [17] selon laquelle on peut considérer la fitness comme tirée aléatoirement d'une distribution de probabilités, Gillespie [18] [19] propose de décrire les mutations bénéfiques en terme de leur statistique en utilisant la statistique des valeurs extrêmes. La statistique des valeurs extrêmes est une théorie mathématique permettant de décrire asymptotiquement le comportement des queues de distribution. L'argument de Gillespie est de dire que si l'on s'intéresse à l'adaptation, alors les fitness que l'on regarde se trouvent loin dans la queue droite de cette hypothétique distribution de probabilités. Une des hypothèses sous-jacentes est que l'on ne regarde que des mutants à une mutations l'individu de départ, aussi appelé wild type. Pour être dans la queue de distribution il hypothétise donc que le wild type est déjà bien adapté. Le sujet sera abordé plus en détail dans la seconde partie de cette introduction.

On peut remarquer qu'il y a un grand espacement temporel entre les travaux de Kimura, Gillespie, Maynard Smith et aujourd'hui. Cette période de quasiment 40 ans n'est bien sûr pas vide de recherches sur l'adaptation et la génétique des populations, et même si mes choix bibliographiques renforcent ce sentiment de vide, cette période marque tout de même un tournant en biologie. Un tournant qui a pu, un temps, faire passer de mode ces questions d'évolutionniste, mais qui permet aujourd'hui de se repencher sur ces questions avec de nouvelles connaissances et moyens expérimentaux. En effet à partir de 1970, on assiste à la montée en force de la biologie moléculaire, et les physiciens n'y sont pas pour rien. Cet essor est notamment dû à une nouvelle compréhension de la molécule d'ADN et de son code, ainsi qu'à un changement de point de vue sur l'expérience. 1953, marque la découverte de la structure de l'ADN par Watson et Cricks [20], rendu notamment possible par les travaux de Chargaff [21] qui en 1952 montre que les rapports entre nombre de bases C/G et A/T valent 1. C'est aussi l'époque du groupe phage dans les années 40. Sous la bannière du biologiste Alfred Hershey, et des deux physiciens Max Delbrück et Salvador Luria (médecin d'origine), le groupe s'intéresse à des modèles simplifiés du vivant tel que les phages. Ils seront aussi à l'origine de la célèbre expérience de Luria et Delbrück [22], qui statura du caractère aléatoire de l'apparition des mutations. 1956 marque la découverte de l'ADN polymérase par Arthur Kornberg, et 1986 signe la première publication publique sur la PCR (Polymerase chain reaction) par Kary Mullis [23]. De 1961 à 1966 le code génétique est décrypté grâce au travail de Nirenberg et Matthaei [24]. En 1965, un modèle pour la régulation génétique est

16. S. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, 128:11–45, 1967

17. M. Kimura. Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl Acad. Sci. USA*, 76:624–626, 1979

18. J. H. Gillespie. A simple stochastic gene substitution model. *Theor. Popul. Biol.*, 23:202–215, 1983

19. J. H. Gillespie. Molecular evolution over the mutational landscape. *Evolution*, 38:1116âĂŞ1129, 1984

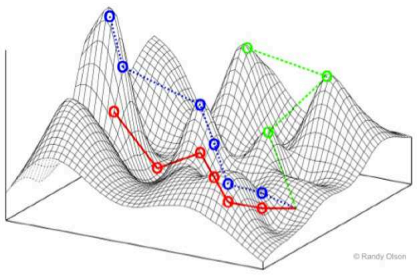


FIGURE 2: **NK modèle en dimension 2 (CC Andy Olson).** Chaque point du plan XY représente un génotype, et suivant Z est encodé la fitness de ce point. Dans ce paysage particulier, qui peut être décrit comme rugueux, plusieurs pics d'adaptation existent. Les traits de couleurs en pointillés représentent quelques chemins possibles dans ce paysage.

20. F. H. C. Crick J. D. Watson. Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953

21. Chargaff E Elson D. On the deoxyribonucleic acid content of sea urchin gametes. *Experientia*, 8:143–145, 1952

22. M. DelbrÃijck S. E. Luria. Mutation of bacteria from virus sensitivity to virus resistance. *Genetics*, 230:491–510, 1943

23. Randall K. Saiki et al. Enzymatic amplification of beta globulin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *JSTOR*, 230:1350–1354, 1985

24. Marshall W. Nirenberg et al. Ribonucleotide composition of the genetic code. *Biochemical and Biophysical Research Communications*, 6:410–414, 1962

proposé par Jacob et Monod avec la découverte de l'opéron lactose . En 1977, Frederick Sanger [25] met au point une technique de séquençage de l'ADN qui permet tout bonnement de lire celle-ci. Ainsi, la biologie se pare de techniques qui lui permettent d'isoler des protéines ou de l'ADN, de modifier des génomes et se concentre alors sur des modèles plus "simple" que l'organisme dans sa totalité, en se focalisant par exemple sur la cellule.

La théorie synthétique de l'évolution comporte des limites déjà largement visibles si l'on veut comprendre l'adaptation au sens large du terme : elle ne prend pas en compte tout ce qui est non génétique comme l'épigénétique (méthylation post transcriptionel, devellopement, effets culturels [26]) ou autres moyens de transmission de l'information/hérédité comme le Lamarckisme [27] ou le transfert horizontal de gènes par exemple. Elle définit pourtant largement le paysage dans lequel cette thèse a baigné et partage avec elle un certain nombre d'intérêts communs. Ainsi le travail de ces trois années a été motivé par les questions suivantes :

- Comment caractériser la sélection ? Peut-on observer des différences dans la manière dont des protéines similaires mais construites différemment se comportent face à la sélection ?
- Des protéines ont-elles pu être designées pour être efficaces face à la sélection ?

Cet étude a été rendu possible par l'énorme boom des techniques de biologie moléculaire et des connaissances sur la cellule et sur l'ADN. Cette thèse se focalise sur la réponse d'un système à la sélection notamment en caractérisant la distribution des plus hautes valeurs de l'observable sous sélection pour le cas d'une famille de protéines, les anticorps. Cette question a ici été abordée expérimentalement lors d'expériences contrôlées, et in vitro, de sélection (phage display) de banques d'anticorps couplées à une analyse statistique quantitative rendue possible par l'utilisation de la technologie de séquençage à haut débit.

ADN polymérase : famille de protéines impliquée dans la réplication de l'ADN et qui assemble les briques élémentaires de l'ADN appelées nucléotides.

PCR : technique de biologie moléculaire qui permet la réplication contrôlée et in vitro d'une partie ciblée d'un brin d'ADN.

Opéron : groupe de gènes qui est contrôlé par un même signal moléculaire régulateur.

25. A. R. Coulson F. Sanger, S. Nicklen. Dna sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, 74:5463–5467, 1977

26. Etienne Danchin and al. Beyond dna: integrating inclusive inheritance into an extended theory of evolution. *Nature Review Genetics*, 12:475–486, 2011

27. Oded Rechavi and al. Transgenerational inheritance of an acquired small rna base based antiviral response in c. elegans. *Cell*, 147:1248–1256, 2011

Lamarkisme : théorie évolutionniste dans laquelle la transmission de caractères acquis par l'expérience de la vie dans un environnent donné est possible.

Transfert horizontal de gènes : capacité qu'ont certaines bactéries, à s'échanger du matériel génétique au cours de leur vie.

Anticorps

Généralités

Un anticorps est une protéine du système immunitaire qui est, tout du moins à une étape de la réponse immunitaire, présentée à la surface des cellules B. Cette protéine présente la spécificité d'évoluer au cours de la réponse immunitaire [28]. En effet, au cours du processus d'évolution appelé maturation d'affinité, la cellule B va subir plusieurs cycle de mutations, de sélection positive, négative et d'amplification. On peut donc supposer que d'une certaine manière ces protéines sont faites pour évoluer.

D'un point de vue strictement structural, un anticorps est composé de deux chaines légères et de deux chaines lourdes Fig.3 . La chaine lourde se compose d'une partie de chaine dite variable et qui est directement la cible de la maturation d'affinité, et d'une partie constante qui permet la transmission du signal de reconnaissance de l'antigène. La chaine légère est plus courte mais est aussi composée d'une partie variable et d'une partie constante. Chez la chaine légère la partie constante est plus courte. Les deux chaines sont reliées par un pont disulfure.

Il y a différentes façons de générer de la diversité dans un anticorps [29]. Tout d'abord il y a de la combinatoire lors de la formation de la région variable. En effet, au moins dans le cas de la formation de la région variable de la chaine lourde, celle-ci est initiée par le réarrangement ADN de trois gènes différents : les gènes V, D et J. Ces trois gènes sont chacun présents sous la forme de plusieurs allèles et un réarrangement VDJ s'obtient par concaténation aléatoire d'un allèle V, D et J. Lorsque les VDJ s'assemblent, des mutations apparaissent dans la zone D, lors de l'accolement de ces 3 gènes. La même chose se produit lors de la formation de la chaine légère mais avec seulement 2 gènes : V et J. Enfin une nouvelle étape de combinatoire est à prendre en compte puisqu'une chaine lourde et légère s'apparient aléatoirement.

La diversité, en plus d'apparaitre dans la combinaison des différents gènes et dans l'aléatoire du réarrangement ADN, apparait aussi au fur et à mesure de la prolifération des cellules B. Des mutations ponctuelles apparaissent alors dans la région variable des deux chaines [30]. Certaines zones de ces chaines sont principalement touchées par ces mutations : ce sont les CDR ou Complementary Determining Region [31]. Ce sont ces CDRs qui sont les sites qui confèrent principa-

28. Simon D. Wagner and Michael S. Neuberger. Somatic hypermutation of immunoglobulin genes. *Annu. Rev. Immunol.*, 14:441–457, 1996

29. Georges A. Gutman. Chapter 8 : genetic basis of antibody diversity. *Medical immunology 544*, Fall 2011

30. Simon D. Wagner and Michael S. Neuberger. Somatic hypermutation of immunoglobulin genes. *Annu. Rev. Immunol.*, 14:441–457, 1996

31. Anderw C. R. Martin Robert M. MacCallum and JAnet M. Thorton. Antibody-antigen interactions : contact analysis and binding site topography. *J. Mol. Biol.*, 262:732–745, 1996

lement la spécificité d'un anticorps. Tout ce qui est dans la région variable et qui n'est pas du CDR, sera dans la suite de cette thèse appelé framework. Une chaine, qu'elle soit lourde ou légère est composée de 3 CDRs. Le CDR3 est considéré comme étant le CDR le plus important pour l'adhésion anticorps/antigène et définit souvent la poche de liaison de l'anticorps. En effet parmi les 3 CDRs c'est lui qui est le plus variable. Il a été montré [32] que le CDR3 de la chaine lourde est plus déterminant que celui de la chaine légère. Ce CDR3 comprend tout le gène D et peut s'étendre légèrement en dehors des interfaces avec les gènes V et J. Les deux autres se trouvent sur la zone V du gène. Il est intéressant de noter que ces CDRs sont la cible privilégiée des mutations, en grande partie parce que des mécanismes d'hypermutations ont été développés à ces endroits, rendant les mutations intrinsèquement non aléatoires : biais dans l'utilisation des codons au niveau des CDRs de sorte qu'une substitution de base entraine un changement d'acide aminé ou encore présence de séquence consensus pour la reconnaissance de l'enzyme AID (qui est enzyme qui induit des mutations de l'ADN).

32. Marc M. Davis Jhon L. Xu. Diversity in the cdr3 region of vh is sufficient for most antibody specificities. *Immunity*, 13:37–45, 2000

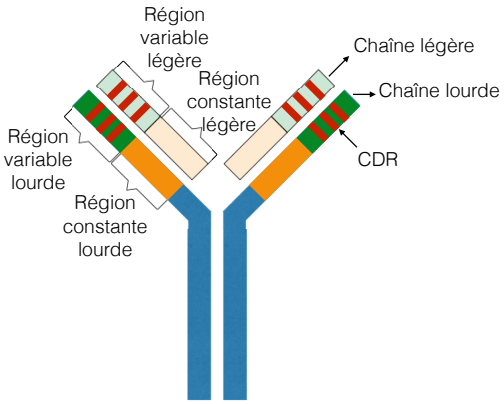


FIGURE 3: **Représentation schématique d'un anticorps.**

On sait déjà que les CDRs sont les groupements de sites d'un anticorps directement responsables de la liaison anticorps/antigène. Pourtant il y a tous ces autres sites entre les différents CDRs, qui définissent le framework et qui comptent pour une part de la variabilité. En effet on peut estimer qu'il existe, codé dans le génome humain, une cinquantaine d'allèles de régions variables V en chaine lourde. Quelle est l'information contenue dans ces régions, qui n'est pas directement une information de liaison ? Y a-t-il une différence en terme d'efficacité à évoluer entre ces différentes régions variables ? Même si ces régions sont moins souvent la cible de mutations, elles sont quand même naturellement mutées. Il est intéressant ici de citer le cas d'anticorps anti HIV [33] qui ont eu leur framework muté à différents taux exceptionnellement haut (jusqu'à 40 pour cent), permettant au malade soit de combattre une souche particulière de HIV soit toutes les souches. Une autre étude, menée par Schultz et al. [34] montre clairement que les mutations qui apparaissent dans le framework jouent un rôle stabilisateur du point de vue de la dynamique. En effet elles viennent contrecarrer des déstabilisations dues à des mutations localisées dans la poche de liaison et qui augmentent l'affinité anticorps/antigène.

33. Florian Klein et al. Somatic mutations of the immunoglobulin framework are generally required for broad and potent hiv-1 neutralization. *Cell*, 153:126–138, 2013

34. Feng Wan et al. Somatic hypermutation maintains antibody thermodynamic stability during affinity maturation. *PNAS*, 110:4261–4266, 2013

Banques d'anticorps dans le contexte de notre expérience

Dans le contexte de l'expérience qui nous intéresse ici, il est important de noter une particularité : on ne présente pas un anticorps entier, mais plutôt des fragments d'anticorps. Ces fragments d'anticorps se composent habituellement tous d'une chaine lourde et légère tronquée pour ne garder que la partie variable. Ces deux fragments peuvent être reliés entre eux de différentes manières (linker

peptidique ou pont disulfure). Ces fragments sont habituellement tirés d'un répertoire immunitaire naturel.
Dans cette thèse les anticorps utilisés sont directement inspirés d'anticorps naturels, bien que modifiés pour permettre une meilleure manipulation de ceux-ci Fig. 4. Enfin, dans les expériences présentées dans ce manuscrit, l'anticorps utilisé se résume à la simple partie variable de la chaine lourde d'un anticorps naturel quelque peu modifié.
Les banques d'anticorps utilisées sont minimalistes et synthétiques. Minimaliste au sens où le fragment de l'anticorps qui est présenté se limite à la partie variable de la chaine lourde et que la diversité n'a été placée que dans les 4 premiers sites du CDR3. Nous avons construit 24 banques d'anticorps qui se différencient par le gène V de la chaine lourde (V_h) utilisée et que nous considérerons à partir de maintenant comme définissant le framework de cet anticorps minimaliste. La plupart de ces banques portent le nom de l'animal à partir duquel son V_h est tiré. Dans toutes les banques le gène J est identique et la diversité est a priori supposé être la même : 4 sites consécutifs dans le CDR3 Fig.4 Fig.6 Fig.7 Fig.8. Les différences entre les banques initiales ont été quantifiées en mesurant la distance entre les différentes banques. En considérant les banques comme une distribution d'acides aminés à chaque sites (sans corrélations : distribution multivariée dont chaque site peut prendre 20 valeurs), on peut calculer la distance, dans l'espace des distributions, qui sépare chaque couple de banque. Dit autrement, on considérera initialement que chaque banque est la donnée de 4 distributions de probabilités, chacune représentant la distribution des 20 acides aminés dans un site du CDR3. Avec cette définition en tête on peut se demander si, étant donné un échantillon d'une banque, on serait capable de reconnaitre de quelle banque cet échantillon provient. Si c'est le cas, c'est très certainement que les banques ne sont pas équivalentes. Ce que l'on veut mesurer ici c'est la différence d'information contenue dans deux échantillons issus de deux banques différentes et si cette différence est significative par rapport à la différence d'information contenue dans deux échantillons venus de la même banque. Pour se faire on peut utiliser une distance basée sur cette quantité que l'on appelle information et qui est le pendant de l'entropie. Cette distance, la distance de Jensen Shannon, est définie comme :

$$JSD(P \,||\, Q) = \frac{1}{2}D(P \,||\, Q) + \frac{1}{2}D(Q \,||\, P) \tag{1}$$

Où P et Q deux distributions et D la divergence de Kullback Leibler définie comme :

$$D(P \,||\, Q) = \sum_i P(i) \log(\frac{P(i)}{Q(i)}) \tag{2}$$

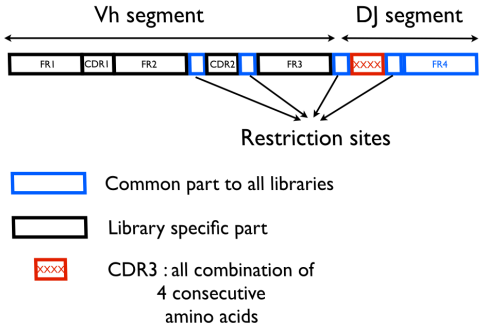


FIGURE 4: **Représentation schématique d'un gène de fragment d'anticorps présent dans nos banques.** Dans notre expérience une banque est définie par le framework des anticorps qui la compose, c'est à dire par le gène Vh de l'anticorps. Toutes les banques ont la même diversité qui se compose de toutes les combinaisons des 4 premiers sites définissant le CDR3 de l'anticorps. Tous les gènes sont flanqués avant le FRW1 d'un site de restriction NCOI, après le FRW4, d'un site de restriction XHOI et entre le FRW3 et le CDR3 d'un site de restriction BSSHII.

Dans notre cas P et Q sont les distributions de CDR3 pour deux banques différentes, où les i se réfèrent aux différentes variables aléatoires. P étant une distribution multivariée de 4 variables aléatoires indépendantes (comme nous le verrons plus tard l'hypothèse d'indépendance est dure à vérifier, et probablement légèrement fausse, au vue du peu de statistique que nous avons eu lors du séquençage des banques initiales) ,$\{X_1, X_2, X_3, X_4\}$ pouvant chacune prendre la valeur d'un des 20 acides aminés (par exemple $x_{10,1}$ représente la réalisation de la variable aléatoire X_1 : acide aminé numéro 10 au site 1) et dont le subscript correspond à un site du CDR3. Il en ressort que dans le cadre de toutes les hypothèses que nous avons faites, la distance entre banques initiales est plus grande que ce à quoi on pourrait s'attendre si on tirait aléatoirement 24 échantillons de la taille de l'échantillonnage de nos données, d'une distribution unique. Fig. 8 et Fig. 9 : en y regardant dans les détails, les différences bien qu'elles soient significatives, ne sont pas non plus outrageusement grandes.
Nous avons restreint la diversité à ces 4 sites, car la technologie de séquençage à haut débit utilisée ne permet de lire que jusqu'à 10^7 séquences : ainsi si on veut pouvoir lire chaque anticorps de la banque au moins 10 fois pour ces 20 banques on peut encore le faire (c'est bien sur sans compter sur le fait que l'on lit plusieurs expériences en parallèle et ces 10^7 doivent être partagés équitablement avec les autres expériences séquencées).
20 des 24 banques sont issues de gènes V_h naturels, bien que modifiés par l'ajout de sites de restrictions, et choisis de sorte qu'ils soient les plus différents possibles Fig.5. Le but ici est d'explorer au maximum l'espace des V_h. Une deuxième contrainte sur ces 20 banques repose sur le fait qu'il a fallu les aligner et que pour cela une séquence de référence est nécessaire. En effet, dans un alignement, il faut savoir à quoi correspondent les sites alignés : on a donc définit les différentes composantes d'un V_h humain, ce qui permet de faire l'hypothèse que les sites qui lui sont alignés représentent les mêmes composantes, typiquement framework et CDRs. Pour un bon alignement les séquences ne doivent pas non plus être trop éloignées de cette séquence de référence.
1 framework est non naturel et correspond à une chaine de glycine. Celui-ci est un contrôle : il n'a pas de structure et n'a pas un historique de sélection.
Les 3 derniers sont des gènes Vh humains : un d'entre eux est un V_h germinal, c'est à dire qu'il n'a pas encore subi de maturation d'affinité, un autre issu du même germinal a subi quelques mutations pour être efficace contre une souche particulière du virus du sida et le dernier, toujours issus du même germinal, a été abondamment muté pour devenir efficace contre un large panel de souches du virus du sida. [35]
Dans cette configuration (frameworks aussi différents que possible et CDR3 aussi similaires que possible), si des différences sont visibles du point de vue de la sélection, alors elles seront en grande partie dues qu'au framework, c'est à dire au choix du gène V_h. Ces 20 V_h

35. Florian Klein et al. Somatic mutations of the immunoglobulin framework are generally required for broad and potent hiv-1 neutralization. *Cell*, 153:126–138, 2013

ont été pris du **V gene heavy chain data base** et correspondent pour la quasi-totalité d'entre eux à des anticorps qui ont déjà subi une maturation d'affinité : à part Nurseshark qui lui est un framework germinal du requin nourrice.

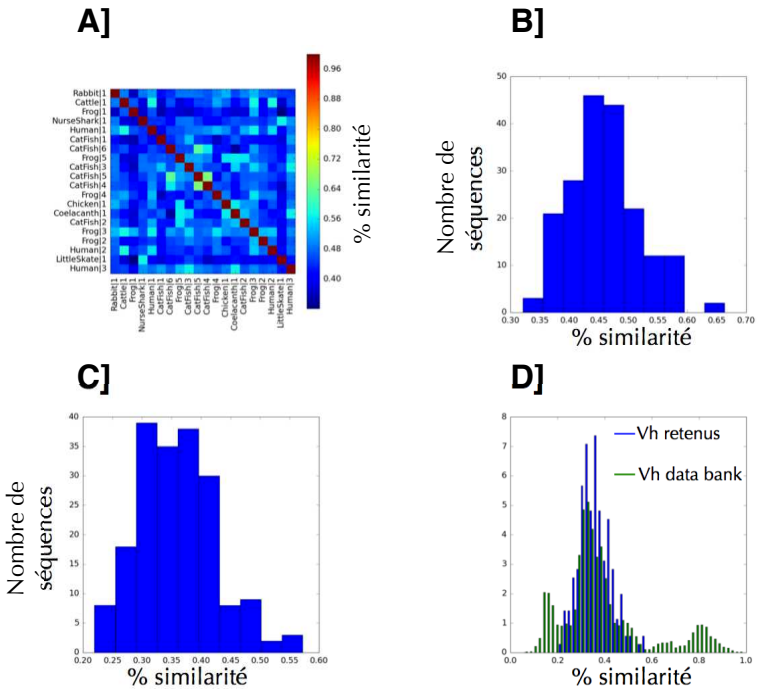


FIGURE 5: **Similarité entre Vh de différentes banques.**Les frameworks des 20 banques ont été alignés.**A]** Matrice de similarité en pourcentage, entre les différents Vh des différentes banques.**B]** Histogramme des similarités entre Vh dans les banques utilisées : ces banques partagent en moyenne 45 pour cent de leur séquences. Il ne faut pas oublier que ces séquences ont été modifiées par l'ajout de sites de restriction communs à toutes les banques.**C]** Histogramme des similarités entre Vh naturel (avant modification par ajout de sites de restriction) dans les banques utilisées : la distribution est cette fois-ci piquée vers les 35 pour cent. **D]** Comparaison de la distribution de similarités entre les Vh utilisés comme bases pour nos banques et la distribution de similarités entre l'ensemble des Vh : les Vh choisis sont effectivement dans la partie gauche, c'est à dire de petit similarité, de la distribution total.

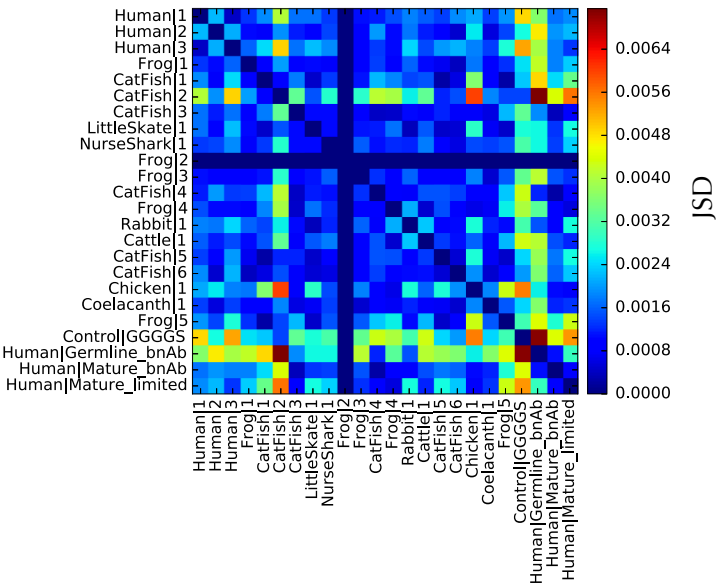


FIGURE 6: **Distance de Jensen-Shannon entre distribution de CDR3 des différentes banques.** Dans l'hypothèse où les CDR3, sont au moins initialement, une séquence de 4 sites totalement indépendants, on peut leur associer une distribution. Celle-ci est définie par la donnée de toutes les probabilités de chaque acide aminé à chaque site : Probability Weight Matrices ou modèle de Pott sans termes d'interaction. La notion de distance existe dans l'espace de probabilité : c'est la distance de Jensen Shannon (JSD). Les JSD entre les distributions de CDR3 ont été calculées à partir des données de séquençage des banques avant sélection. Frog2 est la seule banque que l'on a pas réussi à produire.

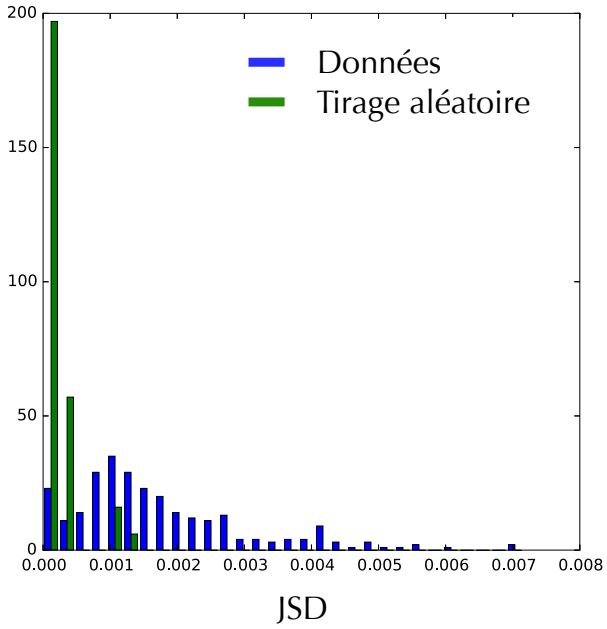


FIGURE 7: **Comparaison entre les JSD de nos banques et les JSD du aux simples effets d'échantillonnage.** Les distances entre les banques calculés précédemment ne veulent rien dire en soi : on ne sait pas si elles sont grandes ou pas car on n'a pas de comparaison. Une des raisons pour lesquelles ses distances ne sont pas nulles est qu'il s'agit de données issues d'un échantillonnage. En prenant toutes les banques ayant la même distribution initiale et en échantillonnant au hasard on peut calculer la distance typique induite entre deux banques, du au simple échantillonnage. Les banques sont plus différentes que ce que le simple échantillonnage peut expliquer.

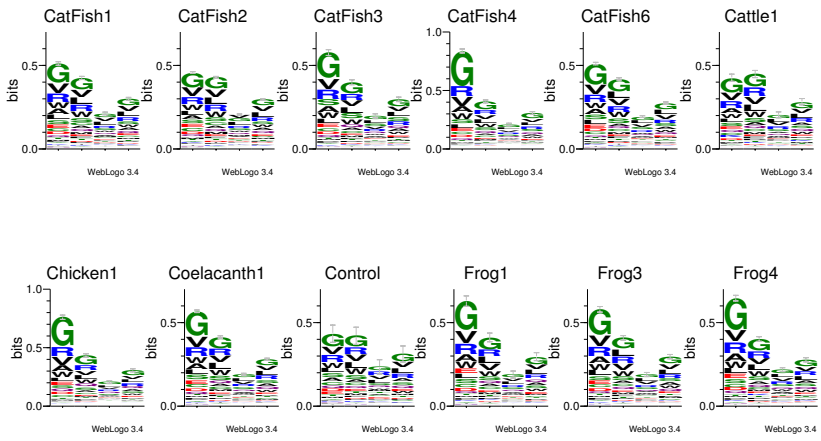


FIGURE 8: **Séquence logo des banques.** Malgré les différences observées en terme de distances, les distributions sont, au moins à l'œil, largement similaires.

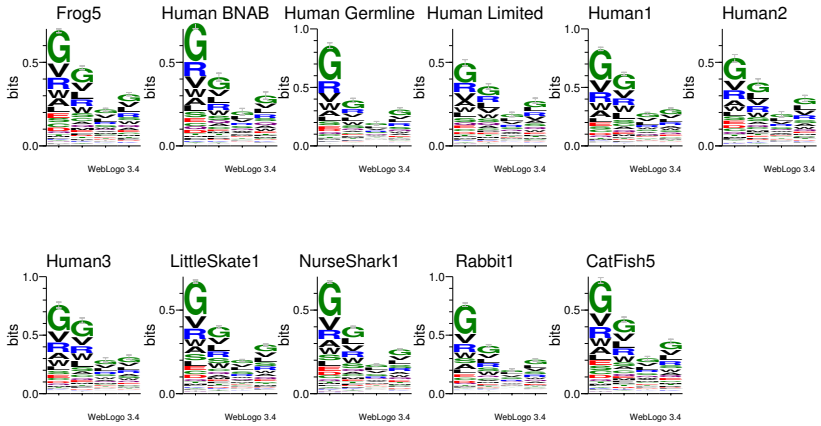


FIGURE 9: **Séquence logo des banques.**

Présentation du phage display

Phage display

Comme présenté précédemment la biologie synthétique a connu un véritable boom dans les années 70. Elle fut mise au service de l'étude de l'évolution et des protéines notamment avec le développement de techniques d'évolution dirigée. Ces techniques d'évolution dirigée permettent de manipuler des populations de mutants et de les soumettre à une évolution contrôlée en reproduisant les trois étapes de l'évolution : mutation, sélection, amplification. Ces techniques d'évolution dirigée se basent sur le principe du Display. Le display, ou présentation, permet de regrouper le phénotype et le génotype d'une protéine. Il est donc possible grâce à ces techniques d'avoir accès à une mesure de phénotype et de pouvoir récupérer le génotype correspondant, pour simplement le lire, le réutiliser ou le modifier. En d'autres termes lorsque le phénotype est sélectionné, le génotype aussi. Pour une revue sur les techniques d'évolution dirigée et de display voir [36]. Les techniques les plus répandues pour faire du display sont le phage display [37],le ribosome display [38], le yeast display [39] et le bacterial display. Ces techniques ne diffèrent pas seulement, comme leur nom l'indique, par le vecteur de présentation de la protéine mais ont chacune leurs avantages et leurs inconvénients. Une des différences principales est la diversité que la population de mutants peut atteindre. Par exemple il est facile de produire 10^{11} phages par ml, mais la concentration maximum d'E.Coli encore en phase exponentielle en milieu riche est de moins de 10^9 bactéries par ml. Ainsi utilisée, une technique de phage display permettra de manipuler jusqu'à 100 fois plus de copies que lors d'un bacterial display. Dans cette thèse nous avons utilisé une technique d'évolution dirigée particulière : le phage display.

Le principe du phage display (tout du moins le phage display que nous avons utilisé), ou présentation d'une protéine à la surface d'un phage est le suivant :

Un phage (virus de bactérie) présente en fusion avec une des protéines formant sa capside, la protéine d'intérêt. Comme dans toutes les techniques d'évolution dirigée, on manipule lorsque l'on fait du phage display, une population de variants de la protéine d'intérêt (équivalent de l'étape de mutation). Il est juste important de réaliser que par phage n'est présent qu'un type de variant de la protéine que l'on veut sélectionner. Ce phage n'est qu'une capside qui

36. Hennie R Hoogenboom. Selecting and screening recombinant antibody libraries. *Nature biotechnology*, 23:1105–1116, 2005

37. G.P. Smith. Filamentous fusion phage : novel expression vectors that display cloned antigens on the vision surface. *Science*, 298:1315–1317, 1985

38. J. et A. Pluckthun Hanes. In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. USA*, 94:4937–4942, 1998

39. M.J. Feldhaus. Flow cytometric isolation of human antibodies from a non immune saccharomyces cerevisiae surface display library. *Proc. Natl. Acad. Sci. USA*, 94:4937–4942, 1998

ne contient aucun de son matériel génétique propre mais seulement une boucle d'ADN, appelée phagemide, et sur laquelle se trouve l'ADN codant pour la même protéine d'intérêt que celle présentée à sa surface Fig.10 .

Si on regarde de plus près, un phagemide est toujours composé de l'ADN codant pour la protéine d'intérêt fusionnée en amont (avant dans le sens de la traduction) de l'ADN codant pour la protéine de surface à laquelle elle sera rattachée. Le phagemide contient aussi un groupe de gènes codant pour le processus d'encapsulation du phagemide à l'intérieur de la capside. Un phagemide a donc sa propre infectivité : il a les moyens de s'encapsuler à l'intérieur d'une capside et ce d'une manière autonome. Souvent le phagemide contient un gène codant pour la résistance à un antibiotique, ce qui permettra plus tard de sélectionner pour la présence du phagemide dans une bactérie infectée.

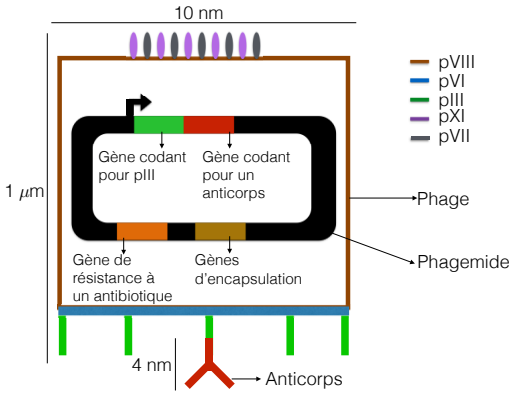


FIGURE 10: **Représentation schématique d'un phage et de son phagemide.**

Puisque la protéine est présentée à la surface, son phénotype peut être mesuré. Dans le cas simple où le phénotype est la capacité à se lier à une cible, la mesure revient à regarder combien de protéines se sont effectivement liées, en comparaison du nombre total de protéines présentées : on a donc accès à la probabilité qu'a cette protéine de se lier. Un simple lavage permet de se débarrasser de la plupart des phages qui ne présentent pas un bon lieur : c'est la phase de sélection. Les phages restants après la sélection sont utilisés pour infecter des bactéries E. Colie. En jouant sur la résistance à l'antibiotique obtenue après l'infection par le phage, on peut s'assurer que l'on ne garde que les bactéries contenant le phagemide et donc infectées. Les bactéries se reproduisent, et avec elles le nombre de phagemide : c'est la phase d'amplification.

La technique de phage display est utilisée principalement pour pouvoir discriminer entre plusieurs mutants (étapes de mutations), les meilleurs pour accomplir une tache : par exemple s'attacher à une cible. Souvent une étape de sélection ne suffit pas à faire apparaitre de grandes différence en terme de nombre entre les différents mutants, et donc à les discriminer, et il faut cycler les étapes de sélection et d'amplification. Il faut donc reproduire des phages à partir des E. Colie infectées. Pour cela un deuxième phage, appelé helper phage intervient en infectant la bactérie contenant déjà le phagemide. Il s'agit d'un phage modifié : il contient toute l'information génétique pour la création d'un nouveau phage mais a une machinerie d'encapsulation mutée de sorte qu'elle soit moins efficace que celle encodé sur le phagemide. Ainsi lorsque toute la capside est formée et qu'il ne reste plus qu'a y incorporer du matériel génétique, c'est le phagemide qui sera encapsulé plutôt que le génome du phage helper.

En pratique, le génome du phage helper contient aussi de quoi coder la protéine de surface qui est fusionnée à la protéine d'intérêt dans le phagemide. Dans le nouveau phage il y aura donc un nombre aléatoire de protéines fusionnées et de protéines de surface non fusionnées. Il est possible de contrôler ce ratio car la production de protéines de surface fusionnées à la protéine d'intérêt est régulée

par un opéron : on peut donc activer ou réprimer sa production. Le génome du phage helper contient aussi un gène de résistance à un antibiotique, différent de celui porté par le phagemide : il est donc possible de sélectionner pour la double infection.
Le phage M13 utilisé en phage display est un phage lysogénique, c'est à dire qu'il ne tue pas la bactérie qu'il infecte pour se reproduire. Il s'agit d'un long filament d'environ un micron de long pour 10nm d'épaisseur. Il est composé de 5 types de protéines de surface : pVII, pIX, pVIII, pVI, pIII. Dans le cas qui nous intéresse la protéine de surface qui présente la protéine d'intérêt est la protéine pIII. Celle-ci est présente en 5 exemplaires en bout de phage et joue un rôle crucial dans l'ancrage du phage sur la bactérie lors de l'infection.
En pratique on peut faire du phage display sans utiliser un phagemide : en fusionnant le gène d'intérêt directement au gène de la protéine de surface qui se trouve dans le génome du phage. Prenons par exemple un anticorps fusionné avec la protéine de surface pIII. Dans ce cas, les 5 protéines pIII seront fusionnées à un anticorps. Le fait de ne pas pouvoir contrôler ce nombre peut être embêtant : en effet on peut se retrouver avec des anticorps qui ont une mauvaise affinité mais qui sont sélectionnés juste parce qu'ils sont 5. On sait aussi peu de choses sur les gênes stériques que peut induire la présentation de ces 5 anticorps. Enfin, l'utilisation d'un phagemide simplifie l'expérience et permet de manipuler plus proprement. En effet, avec le phagemide on ne manipule des phages que lors de la sélection, et pas pendant l'amplification ou le clonage.
Le phage display a été utilisé pour faire évoluer des protéines d'intérêt médicales, pour explorer le fitness landscape de certaines protéines ou encore pour établir les différentes relations entre séquences et fonctions [40].
Dans notre cas bien précis Fig.11, la protéine d'intérêt est l'anticorps, une protéine du système immunitaire.
La technique de phage display d'anticorps date de 1990 [41] quand le groupe de G. Winter montra qu'il était possible de présenter un fragment d'anticorps à la surface d'un phage et que celui-ci gardait suffisamment ces propriétés pour être isolé par sélection dans une population non spécifique. A peine un an plus tard, la première banque (comprenez population de mutants) d'anticorps fut créée et criblée avec succès pour son affinité contre un haptène. 5 ans auparavant il avait été montré qu'il était possible de présenter, un anticorps fonctionnel fusionné à la protéine pIII sans perturber l'infection [42].

40. Douglas m Fowler et al. High-resolution mapping of protein sequence-function relationships. *Nature*, 7:741–746, 2010

41. J. et al. McCafferty. Phage antibodies: filamentous phage displayng antibody variable domains. *Nature*, 348:552–554, 1990

42. G.P. Smith. Filamentous fusion phage : novel expression vectors that display cloned antigens on the vision surface. *Science*, 298:1315–1317, 1985

Notre phage display.

Dans notre cas bien précis Fig.11, la protéine d'intérêt est l'anticorps, un protéine du système immunitaire. Nous travaillons avec des banques d'une diversité de $1.6 * 10^5$. Nous présentons, par expérience, 10^{11} phages à plus de 10^{13} cibles. Cela signifie deux choses

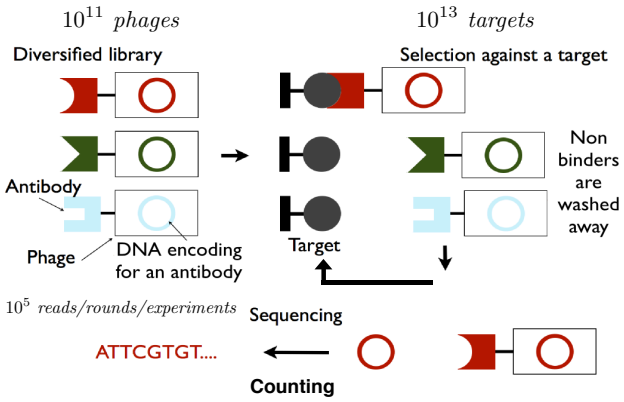


FIGURE 11: **Représentation schématique de notre expérience de phage display.** Une banque d'anticorps fusionnés à la PIII d'un phage est sélectionnée pour la capacité des mutants qui la composent à se lier à une cible donnée. La cible est beaucoup plus présente que les anticorps de sorte qu'il n'y a pas compétition. Après plusieurs tours de sélections suivis d'une étape d'amplification, le résultat de la sélection est séquencé. On a alors accès à l'enrichissement à chaque tours de sélection pour chaque mutant.

importantes :

- Les cibles sont largement en excès par rapport aux phages et il n'y a donc pas compétition entre les anticorps : l'expérience revient à une mesure indépendante de la capacité qu'un anticorps a de s'accrocher à la cible donnée.
- Chaque anticorps particulier est présent entre 10^4 et 10^6 copies : estimation qualitative prenant en compte le biais dans une banque et le bas niveau d'induction de la présentation d'anticorps utilisé.

Les cibles utilisées sont de deux types : PVP et boucle d'ADN [43] [44]. Le PVP a été utilisé au début de nos expériences comme test car il s'agissait d'une cible que nous maitrisions déjà : il s'agit d'un polymère, le Polyvinylpyrrolidone. Cette cible présentant des désavantages de conservation et de chimie de fixation nous nous sommes tournés vers des boucles ADN biotynilés sur des billes magnétiques. Ces boucles ADN se composent d'une tige de 4 bases communes à toutes les cibles et de 9 bases variables qui forment la boucle Fig.12. Ces neufs bases ont été choisies car elles représentent une diversité équivalente aux 4 sites dans nos banques d'anticorps, qu'elles permettent de former une boucle stable et que dans l'hypothèse couramment retenue de l'interaction protéine ADN, permettent une interaction totale de tout le CDR3 avec la boucle. Dans cette banque de cibles, trois cibles ont été isolées pour la suite des expériences : la cible noire, la cible bleue et la cible rouge. Ces trois cibles ont été choisies car ne présentant qu'une seule conformation stable et ne partageant pas deux mêmes couples de bases adjacents. Ces cibles sont biotynilées et fixées sur des billes magnétiques recouvertes de streptavidine.

43. Modi S. et al. Two dna nanomachines map ph changes along intersecting endocytic pathways inside the same cell. *Nature Nanotechnology*, 8:459–467, 2013

44. Soshee A. et al. General in vitro method to analyze the interactions of synthetic polymers with human antibody. *Biomacromolecules*, 15:113–121, 2014

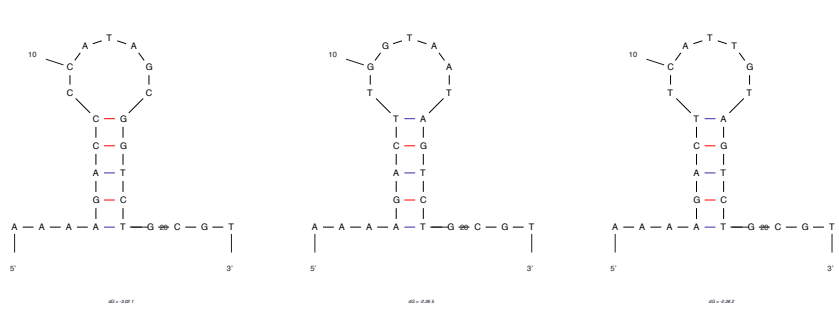


FIGURE 12: **Boucle ADN.** Ces trois boucles d'ADN ont une tige commune et une boucle différente. La tige se compose de 4 paires de bases et la boucle de 9 bases. Les deux boucles sont stables thermodynamiquement et ne présentent qu'une seule structure. Les boucles sont différentes au sens où elles ne partagent pas deux mêmes couples de bases adjacentes.

Nouvelles technologies de séquençage génétique

Présentation

Les nouvelles technologies de séquençage génétique permettent littéralement de lire les bases qui forment une séquence ADN et cela d'une manière massivement parallèle. Il existe un certain nombre de technologies différentes qui se distinguent sur la manière dont sont lues les bases et surtout sur la manière dont les fragments ADN sont immobilisés : fluorescence, détection d'ion, immobilisation dans des nano puits ou sur une plaque de verre traité, etc... Ces différentes technologies ont chacune leurs avantages et leurs inconvénients, notamment sur les longueurs maximales de séquences à lire, le coût de séquençage et les différentes erreurs de lecture.

Nous avons utilisé la technologie illumina : elle permet de lire de 50 à 600 paires de bases avec un nombre total de séquences lues de plusieurs dizaines de millions. La technologie illumina est basée sur la lecture de bases fluorescentes. Les fragments ADN sont munis d'adapteurs qui leur permettront de se lier à la puce de séquençage. Ensuite, lorsque les fragments sont liés une étape d'amplification commence : c'est l'amplification par pont. Lors de cette étape des groupes denses de copies du fragment d'ADN et de son complémentaire sont créés. Des flots des 4 bases fluorescentes sont alors envoyés et se lient ou non d'une façon complémentaire au fragment ADN grâce à l'ajout de primer et d'ADN polymerase. Les bases qui ne sont pas attachées sont lavées. Les bases sont lues par excitation lasers : le signal est assez puissant grâce au groupe dense de fragments.

On peut multiplexer le séquençage, c'est à dire lire plusieurs résultats d'expériences simultanément en confectionnant des codes-barres ADN positionnés avant la séquence d'intérêt.

Enfin, il existe un moyen de quantifier la qualité de lecture de chaque bases : pour chaque lecture de bases un nombre est associé à la base et indique la probabilité d'une erreur de lecture. Ce "quality read" est souvent appelé Q. Une séquence connue est habituellement ajoutée à la lecture pour s'assurer de la corrélation entre ce quality read Q et la probabilité de faire une erreur de lecture. Selon l'appareil, la métrique change légèrement de sorte que l'échelle des Q ne commence pas toujours au même endroit. Pour la technologie que nous avons utilisé, Ilumina MiSeq paire-ended, celui-ci commence à 33. Lorsque l'on a enlevé 33 à toutes les Q on peut alors retrouver la valeur de

l'estimateur pour la probabilité de faire une erreur de lecture.

$$P = 10^{-(Q-33)/10} \tag{3}$$

Lors d'un séquençage, une séquence connue, appelée Phix, est rajoutée pour estimer la qualité du séquençage. Ceci permet aux compagnies de séquençage d'estimer le taux d'erreur lors du séquençage. Le taux d'erreur, calculé à partir des séquences Phix, est habituellement jugé acceptable si il se trouve en dessous de 1.5 % pour des fragments lus de plus de 100 paires de base. Il s'assure aussi qu'un certain pourcentage des bases lues dans ces séquences Phix, ont un Q suffisamment haut, typiquement 30 ce qui correspond à une chance sur 1000 d'avoir mal lu la base.
Toutes ces mesures ne sont que des proxy pour estimer les erreurs de lecture et nous nous attacherons à caractériser les erreurs de lecture dans nos expériences.
Enfin, revenons le temps de quelques phrases, sur la technologie Illumina MiSeq paire-ended. La technologie Mi-seq paire-ended permet de lire la séquence du bout 3' au 5' ainsi que la séquence complémentaire du 5' au 3'. Elle est particulièrement utilisée pour avoir accès à deux lectures indépendantes de la même séquence et donc de réduire la possibilité d'une erreur de lecture. C'est pour cette raison que nous avons choisi cette technologie, pour avoir une double lecture d'une sous partie bien précise du fragment que nous voulions séquencer : le CDR3. Cela n'a pas été possible notamment à cause du design des codes-barres qui ont empêché que la lecture du fragment et de son complémentaire se chevauchent.

Statistiques sur le séquençage

La partie qui nous est inconnue dans le gène de l'anticorps est le CDR3. Celui-ci n'est accessible que dans la lecture du brin complémentaire du fragment. Le CDR3 se trouve dans une zone où la qualité de lecture est en baisse, car celle-ci diminue avec la distance au point de départ de la lecture Fig.13. Pour autant, la grande majorité des séquences se trouvent dans la zone de qualité de lecture acceptable. La lecture de l'autre brin du fragment permet elle, de reconnaitre le framework qui est sélectionné.

On peut estimer la probabilité de faire au moins une erreur sur la lecture des 12 bases formant notre CDR3, en comparant le résultat de lecture de 12 bases jouxtant le CDR3 et dont on connait la séquence Tab.1. Ces erreurs diminuent si on met une contrainte sur la qualité de lecture moyenne de ces 12 bases, mais restent pour certaines expériences non négligeables. Si on estime rapidement le taux d'erreur commis sur la séquence la plus présente au tour 2 ou 3, que l'on prend en compte grossièrement que chaque erreur peut donner lieu à un des 20 acides aminés, l'ordre de grandeurs des séquences

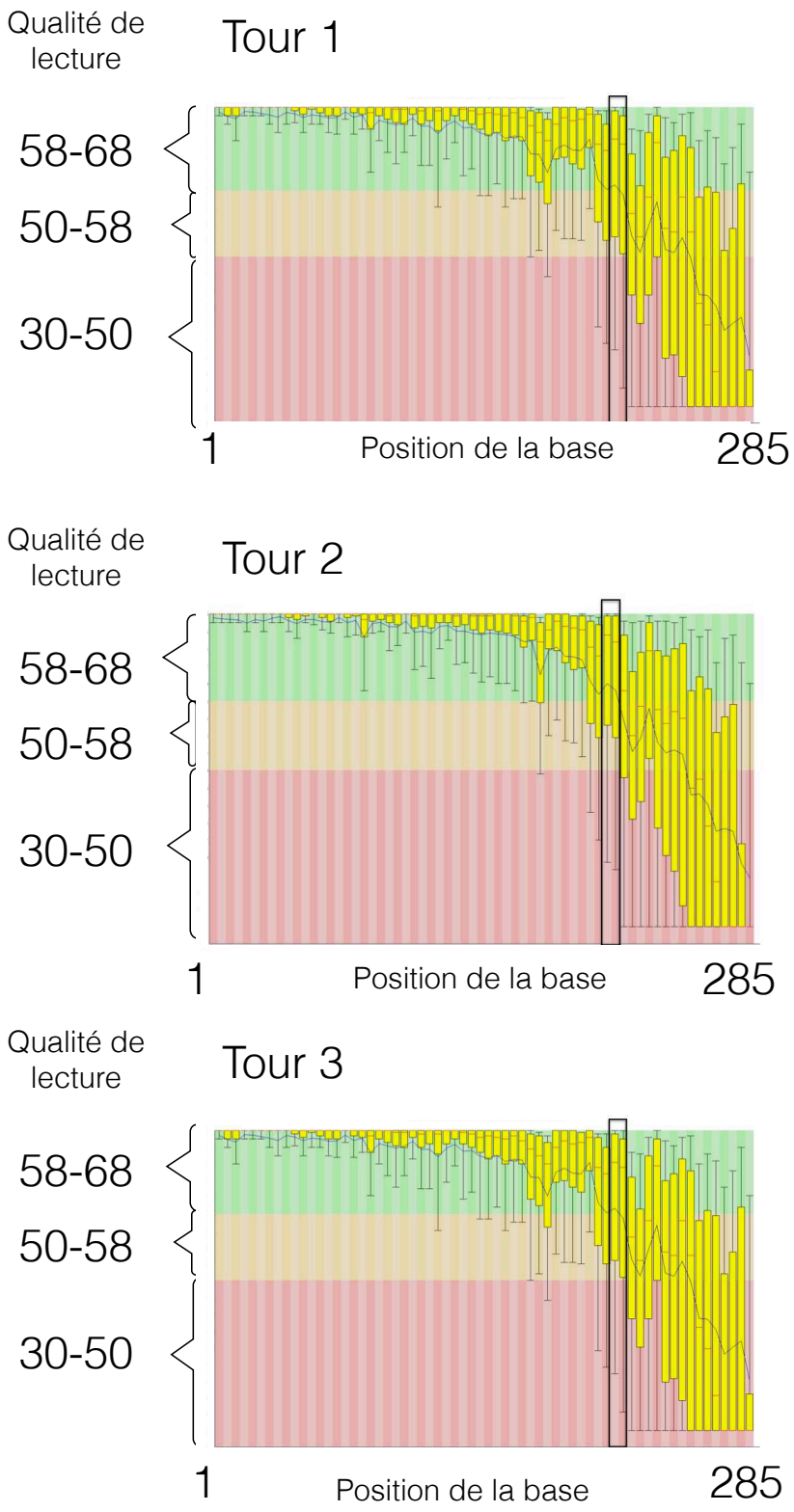


FIGURE 13: **Repartition de la qualité de lecture le long de la séquence lue.** Le séquenceur MiSeq note la qualité de lecture de chaque base lue. Celle-ci à tendance à diminuer lorsque l'on s'éloigne du début de la lecture. La qualité de lecture d'un site sur la séquence lue, est représentée sous la forme de boite à moustache. La ligne grise correspond à la moyenne. La partie verte du graphique représente ce qui est considéré comme une lecture satisfaisante, en orange acceptable, et rouge mauvaise. Sur chaque graphique est représentée une boite noire : ce sont les sites du CDR3. La qualité de lecture sur les sites du CDR3 sont entre acceptable et satisfaisant, même si clairement la majeure partie des séquences sont dans la partie satisfaisante.

qui apparaissent juste par erreur de lecture est de l'ordre de la dizaine de répliques. On pourra donc prendre cela en considération lorsque l'on voudra identifier les séquences que l'on pense valables pour l'analyse.

	Q>30 Fraction de séquences avec au moins une erreur	Q>0 Fraction de séquences avec au moins une erreur
Mix21bis Noire1	0.027	0.104
Mix21bis Noire2	0.046	0.147
Mix21bis Noire3	0.106	0.253
NoShark PVP1	0.025	0.101
NoShark PVP2	0.051	0.160
NoShark PVP3	0.107	0.248
Frog3 PVP1	0.034	0.127
Frog3 PVP2	0.029	0.106
Frog3 PVP3	0.040	0.122
NoFramework Noire1	0.112	0.285
NoFramework Noire2	0.038	0.169
NoFramework Noire3	0.034	0.137
Human2 Noire1	0.043	0.127
Human2 Noire2	0.087	0.221
Human2 Noire3	0.028	0.102
Frog3 Noire1	0.027	0.106
Frog3 Noire2	0.048	0.133
Frog3 Noire3	0.085	0.219
Mix24 Ampli	0.029	0.118
Mix24	0.043	0.125
Mix21bis	0.026	0.102
Mix24 Rouge1	0.030	0.116
Mix24 Rouge2	0.087	0.218
Mix24 Rouge3	0.099	0.186
Mix24 Noire1	0.032	0.119
Mix24 Noire2	0.065	0.177
Mix24 Noire3	0.029	0.117
Duplicate Mix24 Noire3	0.024	0.096
Mix24 Bleu1	0.027	0.106
Mix24_Bleu2	0.049	0.140
Mix24_Bleu3	0.088	0.214
Mix21bis PVP3	0.033	0.138

TABLE 1: **Estimation de $\hat{\kappa}$ pour différents seuil de confiance sur la sélectivité et une qualité de lecture Q>30**

Un récapitulatif de toutes les statistiques sur le séquençage est disponible en annexe.

Réponse d'un système à la sélection

La caractérisation de la sélection reste encore aujourd'hui une question ouverte. Certains se sont attaqués à ce problème en se demandant quelle devrait être la diversité à mettre dans la population de mutants (banque) pour avoir une chance de trouver un certain nombre de bons mutants pour n'importe quelles questions posées (n'importe quelles fonctions voulues). Cette question a d'ailleurs été explicitement posée dans le cadre spécifique de l'étude des anticorps. Expérimentalement il a été montré, dans le cas d'anticorps, qu'une faible diversité pouvait suffire à induire une réponse de sélection contre plusieurs cibles : des banques où les acides aminées des CDRs ne pouvaient contenir que des tyrosines ou des serines ont débouché à la sélection d'anticorps spécifiques contre un facteur de croissance humain [45]. La même équipe s'est intéressée aux effets de la taille de la diversité sur la sélection mais surtout sur la localisation de cette diversité et comment elle devait être introduite : faire tous les mutants à 3 sites, un dans chaque CDR est ce différent de tous les mettre dans un seul CDR du point de vue de la réponse à la sélection [46] ? Une de leur conclusion étant que simplement mettre de la diversité concentrée dans un seul CDR permettait déjà d'avoir des résultats comparables aux banques les plus complexes existantes et ça sans même faire de maturation d'affinité in vitro. Préalablement la question avait été posée par Perelson [47] qui introduit la notion de shape space et avec elle l'idée qu'un anticorps peut s'attacher à plusieurs antigènes de façon spécifique mais avec différentes intensités. D'autres encore, comme cité précédemment se sont intéressés au nombre de maximums dans le fitness landscape propre à un couple protéine/questions posée, ou encore au nombre de chemins qui rendent accessible un maximum si on part d'un wild-type. Enfin il a été montré que même avec une banque totalement aléatoire, c'est à dire pas qui n'est pas construite sur les bases d'une protéine connue, il est possible de trouver de bons mutants [48]. Pour mettre en perspective l'expérience présentée dans cette thèse, rappelons que nous manipulons des banques de diversité 10^5 alors qu'habituellement les banques citées précédemment ont plutôt une diversité comprise entre 10^{10} et 10^9.

45. Frederic A. Fellous et al. Molecular recognition by binary code. *J. Mol. Biol.*, 348:1153–1162, 2005

46. Frederic A. Fellous et al. High-throughput gene ration of synthetic antibodies from bighly functional minimalist phage displayed libraries. *J. Mol. Biol.*, 373:924–940, 2007

47. Alan S. Perelson et Gerard Weisbuch. Immunology for physicists. *Review of modern physics*, 69:1219–1262, 1997

48. Anthony D. Keefe et Jack W. Szostak. Functional proteins from a random-sequence library. *Nature*, 410:715–717, 2001

Valeurs extrêmes

Dans cette thèse je propose de décrire la sélection en terme de valeurs extrêmes. Les valeurs extrêmes sont une branche des statistiques qui s'intéresse à décrire les queues de distributions et se base sur un résultat fort : chaque queue de distributions peut être décrite asymptotiquement par une fonction de probabilité qui se trouve dans l'une des trois classes d'universalités décrivant l'ensemble des queues de distribution. Brièvement, dans le contexte de cette étude, on considère que les différentes capacités à accomplir une tâche des mutants issus d'une banque aléatoire, revient à tirer aléatoirement des capacités d'une distribution de capacités. On s'attellera alors à déterminer si certains mutants sont bien dans la queue de cette distribution, ainsi que la classe d'universalité de la queue.

Suivant le travail de Kimura [49], Gillespie [50] et Orr, nous nous intéressons donc dans cette thèse à caractériser la sélection de différentes banques d'anticorps en utilisant la théorie des Valeurs Extrêmes.

Sous l'impulsion de Orr, l'utilisation des valeurs extrêmes en théorie de l'évolution s'est faite dans un cadre particulier : quelle est la taille des effets bénéfiques de fitness à un point de mutation d'un wild type. Dans ce cas le wild-type défini le seuil. De nombreuses expériences [51] [52] [53] [54] [55] [56] ont été réalisées pour répondre à cette question, à chaque fois sur des organismes complets, comme par exemple des virus, et avec peu de statistiques (une dizaines de mutants bénéfiques, et une centaine tout au plus de mutants). Le résultat de ces expériences montre que l'on retrouve deux des trois classes d'universalité : La Weibull et la Gumbel. Les implications de faire partie d'une ou des autres classes d'universalité sur la dynamique de fixation d'une mutation ont été débattus [57], et montrent clairement que la connaissance de la classe d'universalité est de la plus grande importance : rendant la fixation plus ou moins efficace et de déterministe à aléatoire.

Les thèmes abordés dans cette thèse sont très proches des questions posées par Orr, Kimura et Gillespie, bien que conceptuellement différentes. D'une manière plus rigoureuse que ce qui a déjà été dit précédemment : la théorie des valeurs extrêmes est une branche des statistiques qui s'attache à décrire la distribution d'une variable aléatoire conditionnellement au fait que celle-ci soit supérieure ou égale à un certain seuil. Prenons X, une variable aléatoire décrite par sa fonction de distribution F et u un seuil, alors on s'intéresse en théorie des valeurs extrêmes à caractériser la distribution :

$$\begin{aligned} F^{[u]}(x) &= P(X \leq x \mid X > u) \\ &= P\{X \leq x, X > u\}/P\{X > u\} \\ &= \tfrac{F(x)-F(u)}{1-F(u)}, x \geq u \end{aligned} \tag{4}$$

Le théorème des valeurs extrêmes et plus précisément le théorème de **Balkema-de Haan** permet de décrire asymptotique le comportement de $F^{[u]}(x)$ si x est la réalisation d'une variable aléatoire indépendante

Classe d'universalité : des systèmes très différents aux premiers abords peuvent être décrits par la même statistique à partir du moment où on s'accorde le droit de transformer légèrement la variable aléatoire que l'on regarde. Le théorème de la limite centrale en est un bon exemple. La statistique d'un gaz de Van Der Walls et d'un modèle d'Ising 3D ou la statistique des valeurs propres dans des matrices aléatoires et le spectre d'absorption des atomes, en sont deux autres.

49. Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge Univ. Press, 1983

50. J. H. Gillespie. A simple stochastic gene substitution model. *Theor. Popul. Biol.*, 23:202–215, 1983

51. Angus Buckling R. Craig Maclean. The distribution of fitness effects of beneficial mutations in pseudomonas aeruginosa. *Plos genetics*, 5, 2009

52. Darin R. Rokyta et al. Beneficial fitness effects are not exponential for two viruses. *J. Mol. Evol.*, 2008

53. Rafael Sanjuan et al. The distribution of fitness effects caused by single nucleotide substitutions in an rna virus. *PNAS*, 101:8396–8401, 2004

54. Pedro F. Vale et al. The distribution of mutational fitness effects of phage ϕx174 on different hosts. *Evolution*, 2012

55. Thomas Bataillon Rees Kassen. Distribution of fitness effects among beneficial mutations before selection in experimental populations if bacteria. *Nature genetics*, 2006

56. Darin R. Rokyta et al. An empirical test of the mutational landscape model of adaptation using a single-stranded dna virus. *Nature genetics*, 37:441–444, 2005

57. Paul Joyce et al. A general extreme value theory model for the adaptationn of dna sequences under strong selection and weak mutation. *Genetics*, 180:1627–1643, 2008

X recentrée et rééchelonnée de manière appropriée. Dans ce cas il n'y a que trois classes d'universalité qui permettent de décrire le comportement $F^{[u]}(x)$ dans l'hypothèse où le seuil est suffisamment grand .

$$| F^{[u]}(x) - W_{\kappa,u,\tau_u}(x) | \to 0, u \to sup\{x : F(x) < 1\}$$

$$avec\ W_{\kappa,u,\tau_u}(x) = \begin{cases} 1-(1+\frac{\kappa x}{\tau_u})^{-\frac{1}{\kappa}} & x \geq 0 \quad \text{si} \quad \kappa > 0 \\ 1-(1+\frac{\kappa x}{\tau_u})^{-\frac{1}{\kappa}} & 0 \leq x < \frac{-\tau_u}{\kappa} \quad \text{si} \quad \kappa < 0 \\ 1-e^{-\frac{x}{\tau_u}} & x \geq 0 \quad \text{if} \quad \kappa = 0 \end{cases} \tag{5}$$

Ces trois classes d'universalité sont aussi connues sous le nom de Fréchet ($\kappa > 0$, type loi de puissance), Weibull($\kappa < 0$,type distribution tronquée) et Gumbel($\kappa = 0$, type exponentiel) Fig.14. Le lecteur ayant l'habitude de manipuler les valeurs extrêmes reconnaitra ces noms issus de l'autre pans des valeurs extrêmes : la statistique des maximum. La statistique des extrêmes est plus connue en physique, que la statistique au dessus d'un seuil, mais le lien entre les deux statistiques est directe : on passe d'une distribution des maximums à une distribution au dessus d'un seuil par une simple transformation logarithmique. La convergence de l Eq.5 est une convergence en distribution. $W_{\kappa,u,\tau_u}(x)$ est connue sous le terme *Generalized Pareto Distribution* ou GDP. La densité de probabilité associée est la suivante :

$$w_{\kappa,u,\tau_u}(x) = \begin{cases} \frac{1}{\tau_u}(1+\frac{\kappa x}{\tau_u})^{-\frac{\kappa+1}{\kappa}} & x \geq 0 \quad \text{si} \quad \kappa > 0 \\ \frac{1}{\tau_u}(1+\frac{\kappa x}{\tau_u})^{-\frac{\kappa+1}{\kappa}} & 0 \leq x < \frac{-\tau_u}{\kappa} \quad \text{si} \quad \kappa < 0 \\ \frac{1}{\tau_u}e^{-\frac{x}{\tau_u}} & x \geq 0 \quad \text{if} \quad \kappa = 0 \end{cases} \tag{6}$$

Aujourd'hui, les avancées en biologie synthétique permettent de regarder ce genre de questions au niveau d'une protéine avec tous les avantages que cela implique : simplification du problème (1 gène), diversité et statistique plus grandes. Une des premières explorations que j'ai faite dans le domaine, aura été de prendre les données du groupe Ranganathan [58] et de représenter la distribution des effets de sélectivité Fig.15 pour tous les mutants à un point de mutation du wild-type. La sélectivité pour une protéine dont la fonction est de s'attacher à un ligand, comme c'est le cas dans cette thèse ou dans les données du groupe Ranganathan est juste une mesure de la probabilité qu'a une protéine de passer un tour de sélection. Cela devrait être directement corrélé à sa capacité à s'attacher au ligand contre lequel elle est sélectionnée. Avec cette banque de tous les mutants à un point de mutation du wild type, le groupe Ranganathan à comparer, par rapport au wild type, l'effet de ses mutations sur la sélectivité et cela pour deux cibles différentes.
Il est intéressant de voir comment dès le premier regard les deux queues droites de ces deux distributions ne se ressemblent pas du tout alors que les queues gauches sont plutôt similaires Fig.15. Ma

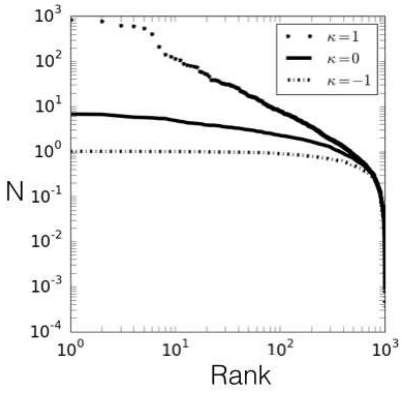


FIGURE 14: **Representation du tirage aléatoire de 1000 variables aléatoires en fonction de leur rang.** Pour chaque classe d'universalité est représenté le résultat du tirage de 1000 variables aléatoires N en fonction de leur rang et d'un paramètre κ particulier. Dans les trois cas le paramètre $\tau = 1$

58. Richard N. MacLaughlin. The spatial architecture of protein function and adaptation. *Nature*, 491:138–142, 2012

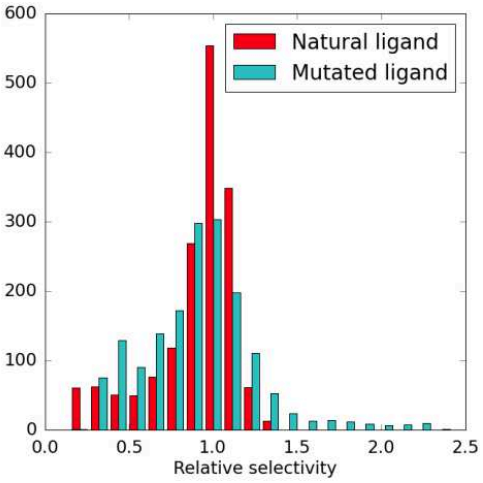


FIGURE 15: **Histogramme de la distribution de sélectivité.** Tous les mutants à un point de mutation d'une protéine wild-type (le PDZ) ont été créés. Pour chacun de ces mutants, leur capacité à s'attacher à un ligand a été mesurée et comparée à la capacité du wild-type. Les différents histogrammes représentent la réponse en terme de capacité à s'attacher à différents ligands : l'un naturel au sens où il est le polypeptide pour lequel le wild type a évolué, et l'autre muté, au sens où il s'agit du ligand naturel légèrement modifié. On peut voir que la réponse aux deux cibles est bien différentes : courte vs longue queue de distribution pour les mutations bénéfiques (sélectivité relative >1). La question est de savoir à quel point ces queues sont différentes.

première question a été, est-ce une vraie différence (c'est à dire que la classe d'universalité n'est pas la même) ? Ou cela résulte-t'il juste de légères différences dans les paramètres décrivant la queue ? A quelle point la réponse est-elle une propriété de la cible ou de la protéine ?

Dans cette thèse je m'intéresse à caractériser une étape de sélection en décrivant les écarts de sélectivité (proxy pour la fitness qui peut et sera définie clairement dans la suite) entre mutants. Si on décide de suivre la vision de Kimura et Gillespie, et considére la sélectivité comme une variable aléatoire, alors décrire les écarts de sélectivité revient à connaitre κ, le paramètre définissant la classe d'universalité. Des simulations Fig.16 montrent que la relation suivante, liant κ et écarts en sélectivité entre deux mutants de rangs consécutifs, semble vérifiée :

$$\log(E[\Delta_i]) = -(\kappa + 1) * \log(i) + \log(E[\Delta_1]) \tag{7}$$

Où $\Delta_i = s_i - s_{i+1}$, E[] la moyenne sur plusieurs tirages, et i le rang. La sélectivité la plus grande est de rang 1, la seconde plus grande de rang 2 et ainsi de suite.

Rang : Le rang est le classement d'un mutant relatif à la valeur de l'observable qui lui est associé : par exemple si on sélectionne pour de l'adhésion on peut classer les mutants dans l'ordre décroissant de leur capacité à s'accrocher à la cible voulue.

Analyse en valeurs extrêmes

La théorie des valeurs extrêmes a été utilisée dans de nombreux autres domaines avant celui de l'évolution et un protocole d'analyse canonique a été développé spécialement dans ce cadre. C'est ce protocole que j'ai utilisé dans l'analyse de mes données de séquençage et que je vais présenter dans cette partie. La plupart des points que j'aborde ici sont tirés de *An introduction to statistical modeling of extreme values* [59], tout du moins en ce qui concerne l'analyse en valeurs extrêmes.

L'observable mesurée est la sélectivité. Elle est habituellement définie comme :

59. Stuart Coles. *An introduction to statistical modeling of extreme values*. Springer, London, 2001

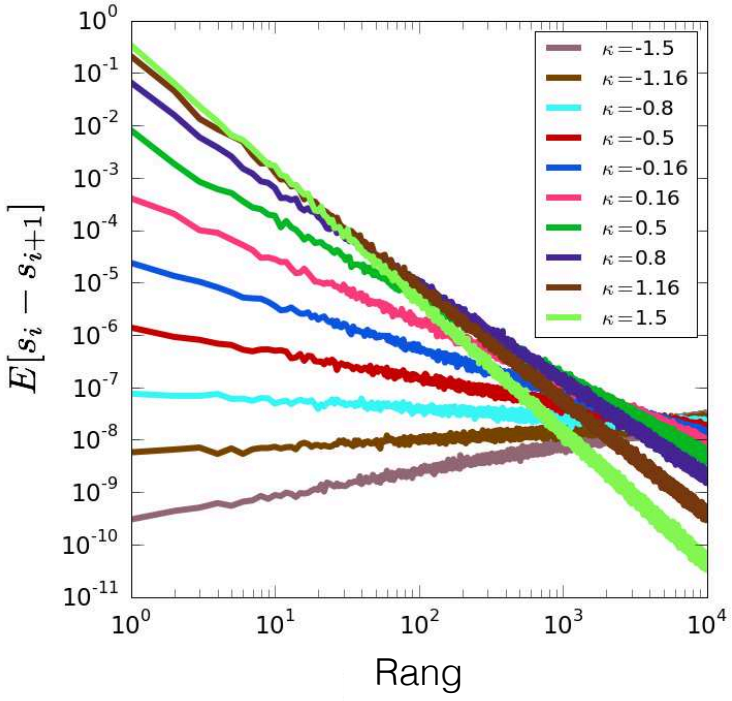


FIGURE 16: **Relation entre le paramètre d'universalité κ et les écarts de sélectivité.** Des simulations montrent que κ est directement relié à l'espacement moyen en terme de sélectivité entre deux mutants adjacents en termes de rang : $\log(E[\Delta_i]) = -(\kappa+1) \times \log(i) + \log(E[\Delta_1])$ où $\Delta_i = s_i - s_{i+1}$, E[] la moyenne sur plusieurs tirages, et i le rang.

$$\hat{s}_j = \frac{n_j^{out}}{n_j^{in}} \tag{8}$$

et représente une estimation de la probabilité s_j qu'a une séquence j d'être sélectionnée. $n_j^{out} et n_j^{in}$ représentent le nombre de séquences j identiques après et avant sélection. Le séquençage ne donne pas directement, accès à cette donnée car entre *in* et *out* une étape d'amplification et d'échantillonnage intervient. On préfère donc définir la sélectivité entre deux tours de sélection comme :

$$\hat{s}_j^t = \frac{1}{Z^t} \frac{\hat{f}_j^{t+1}}{\hat{f}_j^t} \tag{9}$$

Où $\hat{s}_j^t$ est l'estimation de la sélectivité s_j de la séquence j entre les tours t et $t+1$, $\hat{f}_j^{t+1}$ est l'estimation de la fréquence de la séquence j au tour $t+1$ et Z^t une constante multiplicative définit par $Z^t = \sum_{j=1}^{n} \frac{\hat{f}_j^{t+1}}{\hat{f}_j^t}$ introduite de sorte que :

$$\hat{s}_j^t = \hat{s}_j^{t+1} \propto s_j \tag{10}$$

Dans notre expérience, nous mesurons réellement, par l'intermédiaire de cette estimation de la sélectivité $\hat{s}$, la sélectivité intrinsèque de la séquence j : il n'y a pas compétition car il y a au moins cent fois plus

de cibles que d'anticorps.
On va s'intéresser à la distribution de cette variable s. On ne connait pas à priori le lien entre séquence et sélectivité. Une banque peut donc en principe être mappée sur un problème de tirage aléatoire de sélectivités issues d'une certaine distribution de sélectivités. Lorsque l'on opère une étape de sélection on favorise les grandes sélectivités. Si ces sélectivités sont dans la queue de la distribution des sélectivités alors on peut utiliser la statistique des valeurs extrêmes. On sait que le CDR3 est la partie qui a le plus d'effets sur la liaison : on peut donc envisager que les sélectivités mesurables dans une banque de CDR3 correspondent au moins pour un certain nombre, à des sélectivités dans la queue de la distribution de sélectivités.
Pour avoir une première idée sur la distribution on peut représenter ces sélectivités en fonction de leur rang : en partant de l'hypothèse que ces sélectivités ont été tirées uniformément dans la queue alors ce graph revient à présenter la cumulative de la dite distribution Fig.17.
Tout d'abord nous allons ajuster la distribution en utilisant un Maximum Likelihood Estimator (MLE) sur les données, où la fonction likelihood à maximiser est définie comme :

$$\ln L(\kappa, \tau; s_1, ..., s_n) = \sum_{i=1}^{N} \ln f(s_i \mid \kappa, \tau) \tag{11}$$

avec la fonction densité de probabilité paramétrique f étant la fonction densité de probabilité donnée par le GPD Eq.6, et les $\{s_i\}_{0<i\leq N}$ sont les données. Ainsi faut-il faire l'ajustement pour les deux cas : $\kappa \neq 0$ et $\kappa = 0$, qui sont les deux formes du GPD (Eq. 5, montre trois formes mais si on regarde de plus près les formes $\kappa \neq 0$ s'écrivent pareils).Dans un premier temps on fait l'analyse avec la forme du GPD où $\kappa \neq 0$.
La théorie des valeurs extrêmes est une théorie asymptotique qui repose sur le fait que le seuil choisi soit suffisamment grand. Suffisamment signifie ici que l'on se place à un seuil où l'ajustement du paramètre κ est indépendant de ce seuil, tout en gardant le maximum de points. On trace donc l'évolution des différends paramètres de l'ajustement avec les barres d'erreurs associées aux paramètres estimés et on détermine quand le paramètre κ est stable. On s'assure que le paramètre τ suit monotoniquement le choix du seuil à partir du moment où l'on s'approche d'une valeur acceptable pour un seuil, du moins dans les barres d'erreurs. En effet, si on se trouve en présence d'une courbe pouvant être traité avec une GDP, on a la relation suivante est vraie :

$$\tau_{s^*} = \tau_{\hat{s}^*} + \kappa(s^* - \hat{s}^*) \tag{12}$$

Où $\hat{s}^*$ est l'estimateur du paramètre seuil pour lequel on peut estimer que l'on est assez loin dans la queue pour pouvoir être décrit par

une GPD, s^* est le seuil que l'on fait varier pour trouver un plateau sur l'estimateur $\hat{\kappa}$, τ_{s^*} le τ associé au seuil que l'on déplace et $\tau_{\hat{s}^*}$ le τ associé au seuil suffisant pour être dans le cadre des valeurs extrêmes. Plus simplement, lors de l'analyse, on fera varier le seuil s^* jusqu'à ce que cela ne fasse plus varier notre estimateur κ : on trouvera donc $\hat{s}^*$. On s'assurera alors qu'a partir de $\hat{s}^*$ l'estimateur τ_{s^*} soit bien linéairement dépendant de s^* avec une pente κ.
Les estimateurs pour les différents paramètres sont munis de barres d'erreurs. Le calcul des barres d'erreur se fait suivant l'hypothèse de normalité de l'estimateur, ici pour l'exemple τ_{exp} qui est l'estimateur dans le cas exponentiel :

$$\sqrt{n}(\hat{\tau}_{exp,n} - \tau_{exp}) \xrightarrow{D} Normal(0, I(\tau_{exp})^{-1}) \tag{13}$$

Où I est l'information de Fisher et D désigne la convergence en terme de distribution. De part la forme de Eq. 13, on peut apparenter l'inverse de cette Information de Fisher à une variance. Dans le cas multivarié (nous avons en effet deux paramètres à ajuster simultanément) qui nous intéresse il faut inverser la matrice d'information de Fisher définie ainsi :

$$I(\kappa, \tau)_{i,j} = E[(\frac{\partial}{\partial \kappa} \ln L(\kappa, \tau; S))(\frac{\partial}{\partial \tau} \ln L(\kappa, \tau; S))] \tag{14}$$

Où $E[]$ est la moyenne sur le tirage et S notre variable aléatoire. La racine carrée des termes diagonaux de la matrice inversée est prise pour déviation standard et multipliée par 1.96 pour donner l'intervalle de confiance à 95 pour-cent autour des estimateurs. Sous la forme de Eq.14, on voit que l'information de Fisher correspond à la donnée des dérivées secondes dans l'espace des likelihoods. Si on en prend l'inverse, cette I^{-1} est assimilable à la courbure dans l'espace des likelihoods aux alentours d'un maximum. En d'autres termes I^{-1} nous dit à quel point le maximum qui nous a permit de trouver $\hat{\kappa}$ et $\hat{\tau}$ est plat ou pentu. Plus il est plat, moins on a de résolution sur la véritable valeur de nos estimateurs et plus les barres d'erreurs associées sont grandes.

Un Maximum Likelihood Estimator donnera toujours un résultat même si l'apriori sur la fonction à ajuster est mauvais. Donc lorsque les paramètres ont été estimés on s'assure que le MLE donne effectivement un bon ajustement sur les données. Pour cela, de par la nature même des valeurs extrêmes il est coutume de vérifier la plausibilité de l'ajustement sur deux types de courbes : la fonction cumulative et la fonction quantile associée à notre variable aléatoire sélectivité. Après une définition de ces deux fonctions, nous verrons pourquoi dans le cas des valeurs extrêmes ces deux fonctions apportent des informations complémentaires. On notera au passage que l'ajustement s'est fait sur la fonction densité et donc que cette

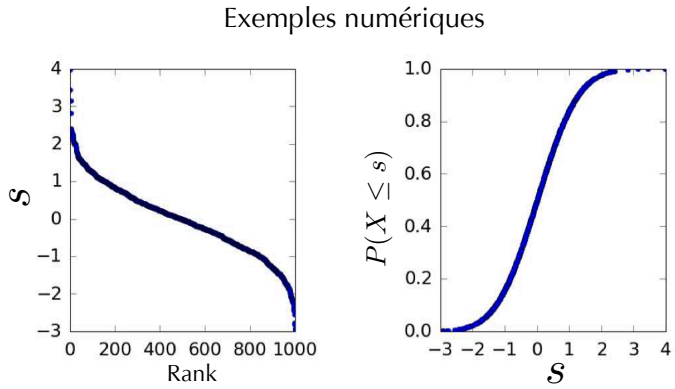


FIGURE 17: **Relation entre rang vs sélectivité et cumulative.** On peut définir le rang de la sélectivité s par $r(s)$ et définir $P(X \leq s) = 1 - \frac{r(s)}{D}$ ou D est la diversité.

vérification de plausibilité s'assure que l'ajustement contient des informations avec d'autres types de statistiques que celle directement utilisée pour l'ajustement. Le cumulative est définie comme :

$$F_S(s) = P(\mathbf{S} \leq s) \tag{15}$$

et représente donc la probabilité qu'une réalisation de **S** soit inférieure à s : à une réalisation de **S** est donc associée une probabilité. Alors que la fonction quantile elle est définie par :

$$Q(p) = \inf\{s \in \Omega : p \leq F_S(s)\} \tag{16}$$

(Ω est l'ensemble des réalisations de **S**) et représente donc la réalisation de **S** pour laquelle p% des réalisations sont plus petites que s : à une probabilité est donc associée une réalisation. Donc dans les comparaisons de cumulatives, des probabilités seront comparées, et dans le cas de comparaisons de fonctions quantile se sera des réalisations de **S** qui seront comparées. Ces deux comparaisons permettent de magnifier différemment les points expérimentaux. La fonction quantile revenant à prendre l'inverse de la cumulative, on se retrouve à tirer des variables aléatoires : ainsi le graph accentue la queue de la queue. Pour la cumulative tous les points sont homogénement représentés. On a donc pour le cas de la fonction quantile, une représentation des s issus des données vs s issus du modèle où on est sensible aux grands s (ceux qui nous intéressent par dessus tout) mais quasiment aveugle s plus petits. Avec la fonction cumulative on regarde la qualité de l'ajustement d'une façon plus globale. Dans les deux cas, données et points du modèle doivent se retrouver sur une ligne : $y = x$ dans le cas de la cumulative, $y = \hat{s}^* + \tau s$ dans le cas de la fonction quantiles. Ainsi pour la fonction quantile on a besoin de ne poser que κ et le graph permet de retrouver directement τ et le seuil $\hat{s}*$.

En plus d'une simple visualisation de l'ajustement d'une des deux formes du GPD, on peut estimer à l'aide d'une P-value (P comme

Pearson, très brièvement cité dans l'introduction) la probabilité de ne pas pouvoir différencier une distribution de Gumbel (exponentielle) et une des deux autres classes d'universalité définies par deux paramètres (Fréchet et Weibull). En effet, il faut noter que le GPD a deux formes non équivalentes : une à deux paramètres κ et τ et l'autre à un seul paramètre τ_{exp}. Plus on ajoute de paramètres meilleur sera l'ajustement. Mais cela a un coût que l'on n'est pas forcément prêt à payer (compléxification du modèle pas forcément nécessaire, perte de signifiance du modèle...) et on préférera ajuster avec le moins de paramètres possibles si nos données ne contiennent pas assez d'information pour faire la différence entre les deux formes du GPD d'une manière significative. On définit alors un likelihood ratio test (On aurait aussi très bien pu utiliser une distance de Jensen Shannon et on verra que c'est quelque part ce que l'on fait, mais une autre mesure bien plus adaptée à la signifiance d'un ajustement existe), noté LRT, qui donne une différence de score relatif à la pertinence de l'estimateur entre les deux distributions, relativement à nos données. On définit de nouveau la fonction log-likelihood :

$$l(\mathbf{S} \mid \hat{\kappa}, \hat{\tau}) = \sum_{i=1}^{N} \ln(f_{\mathbf{S}}(s_i \mid \hat{\kappa}, \hat{\tau})) \tag{17}$$

Où nos données sont $\{s_1, s_2, ..., s_N\}$ et $f_{\mathbf{S}}(s_i \mid \hat{\kappa}, \hat{\tau})$ la valeur du GPD pour cette valeur expérimentale et pour ce jeu de paramètres estimés. A partir de cette formule on définit une variable aléatoire, la LRT :

$$-2\ln(\Lambda) = 2 * (l(\mathbf{S} \mid \hat{\kappa}, \hat{\tau}) - l(\mathbf{S} \mid 0, \hat{\tau}_{exp})) \tag{18}$$

Qui permet d'établir une différence de score entre les deux formes du GPD à partir de nos données. Enfin on regarde comment est distribuée la valeur du LRT (en simulant la valeur d'un LRT sur 10 000 tirages de N (la taille de l'échantillon que l'on ajuste) variables aléatoires issues d'une exponentielle avec le bon τ_{exp}) et on compare avec la valeur obtenue pour nos données : on a ainsi accès à une P-value qui nous indique le pourcentage de chance que l'on aurait de prendre la forme exponentielle (Gumbel) du GPD pour son autre forme (Fréchet ou Weibull). En effet, on mesure ici à quel point une telle différence de score entre ces deux distributions est habituelle. On considérera une P-value inférieure à 10^{-4} comme bonne et inférieure à 10^{-3} comme acceptable.
Ce paragraphe un petit peu technique (et non essentiel à la compréhension des résultats exposés par la suite) s'adresse avant tout aux lecteurs intéressés par la théorie de l'information et toutes les mesures que j'ai introduite depuis le début de ce chapitre. La forme du LRT telle qu'il est écrit Eq.18 peut paraitre un peu déconcertante : pourquoi garder ce 2 par exemple ? La raison est en faite historique : Samule S. Wilks a montré que la statistique du LRT sous cette forme tendait asymptotiquement pour des tailles d'échantillonnage infinies

vers une distribution χ^2. Ce 2 est probablement là pour pondérer le LRT avec la variable aléatoire **S** comme le ferait une distribution χ^2. En effet la somme de variables aléatoires indépendantes et normales mises au carré, peut définir une nouvelle variable aléatoire et celle-ci est distribuée selon un χ^2. On ne connait malheureusement pas la taille de l'échantillonnage pour que ce χ^2 soit une bonne approximation de la statistique du LRT et que l'on puisse en déduire une P-value. C'est pour cela que l'on procède à une simulation pour obtenir une distribution approchée du LRT et ainsi calculer la P-value associée. En réalité, pour la plupart des tailles d'échantillonnage pour lesquelles nous avons ajusté les données, la distribution du LRT convergeait fortement vers un χ^2 et la P-value associée était du même ordre de grandeur que celle simulée. Dans le même ordre d'idées, j'ai mentionné que l'on aurait pu utiliser une distance de Jensen Shannon au lieu de ce Likelihood Ratio Test. En faite, pour des tailles d'échantillonnage infinies, le LRT entre modèle nul (τ_{exp}) et modèle à deux paramètres (κ et τ) revient à calculer une divergence de Kullback Liebler pour ces deux modèles et notre jeu de données : il suffit pour cela de dire que la loi des grands nombres autorise à approximer le LRT par sa moyenne . On retombe ainsi naturellement sur la définition de la distance de Kullback Liebler.

Analyse générale

En dehors de l'analyse en terme de valeurs extrêmes, on peut faire une analyse plus conventionnelle, et qui consiste à regarder la distribution des acides aminés en fonction des sites du CDR3, pour déterminer l'apparition d'une séquence consensus. On peut aussi regarder les corrélations entre ces sites. Il s'agit certes d'un type d'analyse plus commun qu'une analyse en valeurs extrêmes, mais elle est intimement liée à la vision que nous avons des valeurs extrêmes. En effet jusqu'à maintenant je n'ai parlé que de sélectivité qui peut être considérée comme une observable macroscopique. En regardant les fréquences d'acides aminés par site et les corrélations entre sites, j'introduis une notion de variable microscopique, le site, et dont la distribution est potentiellement liée à la distribution des sélectivités. C'est une information dont on dispose grâce au séquençage mais jusqu'à présent non utilisée. En effet, une étude postérieure d'un répertoire immunitaire naturel [60] a montré qu'il existait un lien entre corrélations en terme de sites de séquences d'anticorps et l'observation de loi de puissance dans la distribution d'anticorps dans un répertoire immunitaire. Il est donc intéressant de voir si l'on retrouve de telles lois dans nos résultats et si l'on peut nous aussi observer des corrélations entre sites du CDR3 lors de la sélection.

60. William Bialek Thierry Mora, Aleksandra M. Walczak and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA*, 107:5405–5410, 2010

On mesurera la distribution des acides aminées en fonction des sites du CDR3 en calculant l'entropie relative, aussi appelé divergence de Kullback Liebler, en bit sur un site k. Ces entropies sont relatives à la distribution initiale, potentiellement biaisée, des acides aminés. Ces données seront représentées sous forme matricielle, avec le détail de

la contribution en terme d'information de chaque acide aminé sur chaque site. Dans ce cas on définit plus précisément l'entropie relative comme :

$$D(X_k \,||\, Y_k) = \sum_{i \in aa} P(x_{i,k}) \log_2(\frac{P(x_{i,k})}{P(y_{i,k})}) \tag{19}$$

Où k correspond au site que l'on regarde, aa l'ensemble des acides aminés, X_k la distribution des acides aminés à un tour donné de sélection pour le site k et Y_k a distribution des acides aminés dans la banque initiale pour le site k. l'utilisation de l'entropie relative est standard : c'est cette mesure que l'on retrouve dans les séquences logo Fig.8. Si sélection il y a, un sous-groupe d'acides aminés devrait voir leur entropie relative croitre avec les tours de sélection. D'un point de vue pratique cette distribution sera présentée sous forme matricielle : chaque ligne représentant la contribution de chaque acide aminé en terme d'entropie relative par site. Il est donc possible d'avoir des contributions négatives de certains acides aminés à l'entropie relative d'un site, même si au total bien sûr l'entropie relative reste positive.
Pour regarder les corrélations entre sites on décide de mesurer l'information mutuelle. Celle-ci est définie de la manière suivante :

$$MI(X_k, X_l) = \sum_{i \in aa} \sum_{j \in aa} P(x_{i,k}, x_{j,l}) \log_2(\frac{P(x_{i,k}, x_{j,l})}{P(x_{i,k})P(x_{j,l})}) \tag{20}$$

Sous cette forme il est clair que l'information mutuelle est une mesure de la corrélation entre deux variables aléatoires qui sont ici nos sites. Lorsqu'il n'y a pas corrélation alors $\frac{P(x_{i,k}, x_{j,l})}{P(x_{i,k})P(x_{j,l})} = 1$ et l'information mutuelle vaut 0. Toutes déviations de ce 0 signifient la présence de corrélation entre les sites k et l.

Résultats

Dans cette partie nous présentons les résultats des expériences de sélection de nos banques minimalistes et synthétiques. Réussir à sélectionner d'une façon reproductible et spécifique des anticorps aussi synthétiques à partir d'une banque d'aussi faible diversité et contre des cibles diverses, est en soi un résultat fort.
Une analyse plus habituelle sur les résultats de la sélection en terme de séquence consensus et de corrélation entre sites du CDR3 sera tout d'abord entreprise, et complétée par une analyse de la distribution de sélectivité dans ces expériences.

Présentation des mixtures de banques

Nous avons mixé différentes banques ensemble, et procédé à un crible contre différentes cibles. Des mixtures différentes existent (soit toutes les banques, soit certaines absentes...) et ont pour but de pouvoir comparer les banques entre elles.. Une première banque, Mix 21, est sensée être un mélange homogène de toutes les banques, sauf des banques issues des frameworks VIH Fig.18. L'homogénéité pouvant être améliorée, une deuxième banque Mix21 bis a été refaite plus précautionneusement. Une mixture de toutes les banques a aussi été faite : Mix 24 Fig.19. Nous en avons profité pour nous assurer que l'étape d'amplification (production de phage en milieu E.Coli saturée et infection par ces phages d'E. Colie fraîches) ne favorisait pas une banque. Il se pourrait en effet, qu'un framework empêche la bonne infection de E. Colie par le phage, ou qu'un phagemide soit significativement moins couteux à répliquer pour la bactérie et donc qu'elle pousse plus vite et envahisse la culture. L'amplification ne semble pas faire de différences importantes entre les banques Fig.19.
Enfin un troisième type de banque a été testé : la NoShark. Elle avait pour but d'être une Mix24 sans NurseShark, mais pour des raisons de défauts d'étiquetage, cela n'a pas été le cas.
Dans la majeure partie des cas ces mixtures voient sélectionner un seul framework.

Reproductibilité de la sélection

Deux expériences ont été répliquées pour s'assurer de la reproductibilité de la sélection. Seul les tours 3 de ces répliques ont été séquencés et comparés aux tours 3 dits "originaux". Cette comparai-

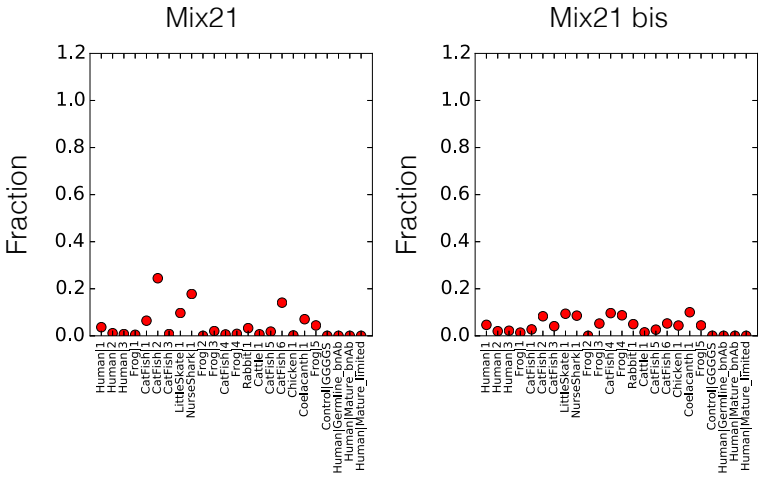


FIGURE 18: **Composition de Mix21.** Dans Mix21, des écarts importants sur l'homogénéité de la mixture sont visibles. Pour Mix21bis, on retrouve la fraction attendue de 0.05 à 0.1 de la mixture occupée par chaque banque.

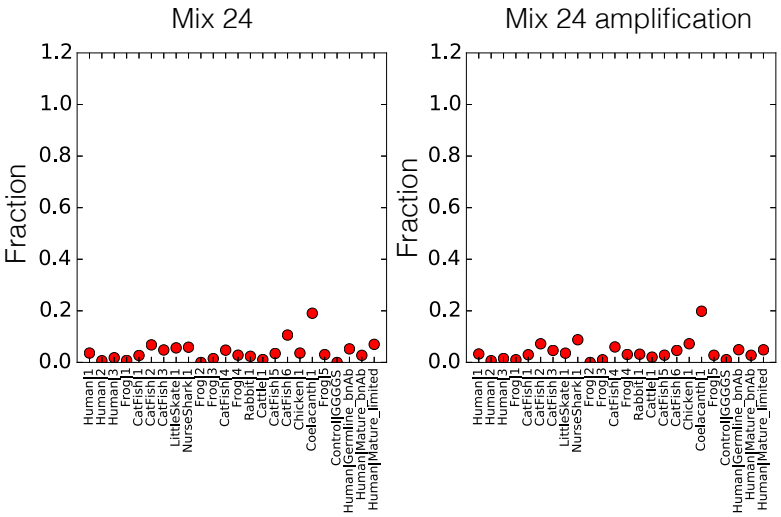


FIGURE 19: **Composition de Mix24** Aucune différence significative n'est observable entre la mixture initiale et la mixture initiale ayant subi une étape d'amplification sans sélection.

son se fait en terme de fréquences et de rangs pour l'expérience dite de Mix24 contre noire et de Mix21bis contre PVP. Le résultat de la sélection après le troisième tour de sélection est comparé entre original et réplique, en terme de rang et de fréquence d'apparition Fig.20. En terme de fréquences ou de rangs, on se rend compte que les expériences sur la cible ADN ou PVP sont plutôt reproductibles. Le fait que l'expérience contre la cible noire donne une corrélation plus grande entre les fréquences des séquences partagées par les deux tours 3 en comparaison de l'expérience contre PVP est attendu. En effet, les écarts entre fréquences étant plus grand pour l'expérience contre noire, on peut s'attendre à ce qu'il y ait moins cet effet nuage de points. Il est cependant à noter que pour la cible PVP il est à un biais remarquable au niveau rang (on le voit aussi en fréquences mais c'est moins net) : le biais a tendance à décaler les séquences nombreuses (petit rang) dans l'expérience originale vers les hauts rangs dans la réplique.

En terme de fréquences et de rang, on peut dire que l'expérience

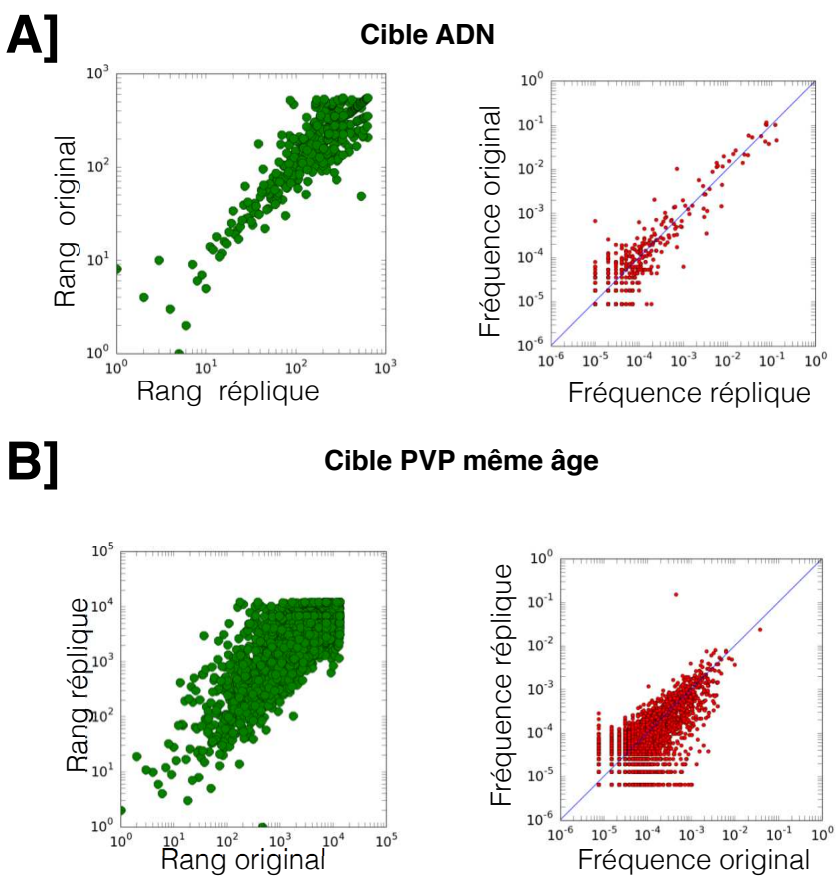


FIGURE 20: **Comparaison des résultats de sélection au tour 3. A]** Comparaison pour une sélection contre la cible noire. La sélection est plutôt reproductible même si il peut y avoir des écarts de l'ordre de 10 en terme de rang même pour les rangs les plus hauts. **B]**Comparaison pour une sélection contre la cible PVP. La sélection est moins reproductible que pour la cible noire, mais cela peut aussi être dû à la différence en terme de nombre de séquences partagés entre les deux expériences, ce qui est équivalent à dire que les fréquences en **A]** sont plus espacés que dans cette expérience : il y a un effet nuage de points .On remarque aussi que la séquence au rang 1 dans la réplique est 400éme dans l'expérience originale et que de nombreux hauts rangs dans l'expérience originale se retrouvent relégués à des rangs beaucoup plus bas dans la réplique.

est plutôt reproductible pour l'expérience contre PVP et reproductible pour l'expérience contre noire.
Comme nous allons le voir dans la section suivante la reproductibi-

lité de la sélection peut même être étendue à une analyse encore plus fine.

Résultats des sélections en terme de séquences

Les données présentées dans cette section ont subi un premier tri : seules les séquences dont les CDR3s présentent en moyenne une qualité de lecture Q>30 (63 selon la classification Illumina voire chapitre *Séquençage*) ont été gardées. L'entropie relative ainsi que l'information mutuelle de chaque site du CDR3 a été calculée quand cela a été possible, pour les 3 tours de sélection qu'ont subi chaque expérience. Cette analyse permet de mettre en évidence des séquences consensus stables au cours des tours, spécifique de la cible, mais aussi spécifique du framework dans lequel le CDR3 interagit. Pour faire le lien avec la partie *Reproductibilité de la sélection*, intéressons-nous tout d'abord aux deux expériences suivantes.

MIx21 et NoShark contre PVP

Nous avons deux banques qui donnent le même résultat contre PVP : la Mix 21 et la NoShark. Pour le rappel, NoShark est la banque Mix24 sans Frog3 bien qu'à l'origine nous voulions qu'elle soit sans NurseShark. Avant d'aller plus loin, il existe une autre différence entre les résultats de l'expérience Mix21 contre PVP et NoShark contre PVP : Mix21 contre PVP a été séquencée à 10^6 fragments lus en moyenne par tour, alors que NoShark contre PVP, seulement à $2 * 10^5$ fragments lus par tour (ce qui est la moyenne pour toutes les autres expériences présentées). Ces deux banques criblées contre PVP, à quasiment un an d'intervalle, montrent un pouvoir de sélection et de spécificité pour le framework NurseShark Fig.21 et Fig.22 et une même séquence consensus "GWYT" Fig.**??** et Fig.24.Cette séquence consensus évolue "continument" avec les tours. La reproductibilité de la sélection de NurseShark contre PVP se retrouve aussi dans les corrélations entre sites Fig.26 et Fig.27.
De plus on voit très bien les corrélations évoluer au cours des tours : partant d'une distribution de corrélations très proche de celle présente initialement dans la banque NurseShark pré sélection Fig.25 , pour arriver au tour 3 à une distribution totalement différente. Le tour 2 fait parfaitement la jonction entre les deux distributions. C'est encore là une démonstration de la nécessité de faire au moins 3 tours de sélection pour voir du signal de sélection émerger.
Toutes les corrélations présentées ici sont comparées aux corrélations induites par un modèle sans corrélation, respectant les fréquences de chaque acides aminés à chaque sites et échantillonnée de la même façon que dans les données. Toutes les corrélations sont plusieurs ordres de grandeurs au-dessus des corrélations dues à l'échantillonnage et sont donc significatives.
Dans le cas de la Fig.27 cela signifie que les corrélations observées sont bien le fruit d'une sélection. Dans cas de Fig.25, elles signifient

que l'hypothèse de totale indépendance des sites est fausse. Le fait est que pour toutes les autres banques, et parce qu'elles ont été moins séquencées, il est impossible de réfuter l'hypothèse d'indépendance. Pour autant, comme nous allons le voir, la forme des corrélations initiales entre sites pour la totalité des banques initiales est très similaire à celle de NurseShark initiale. On peut faire l'hypothèse à partir de là que toutes les banques (forte extrapolation) ont initialement des corrélations entre sites, suivant celles de NusrseShark. Pour autant comme on n'a pas vraiment de moyen de le vérifier, jusqu'à séquençage plus dense, les notions de distances présentées dans la partie *Banques*, restent le travail le plus poussé accessible.

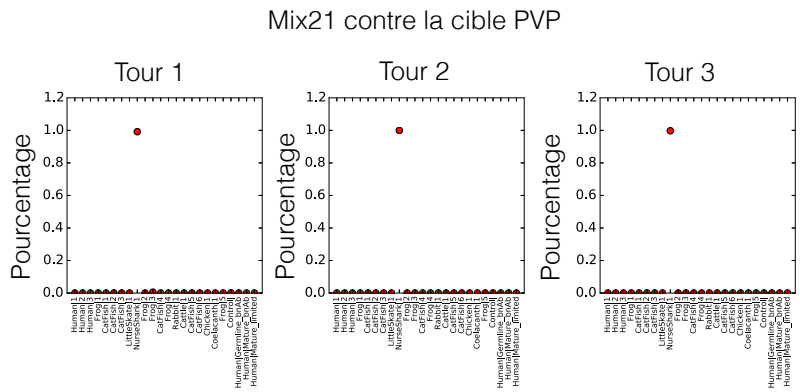


FIGURE 21: **Répartition des frameworks dans Mix 21 avec les tours de sélection contre PVP.** Dès le premier tour NurseShark est très majoritairement sélectionné.

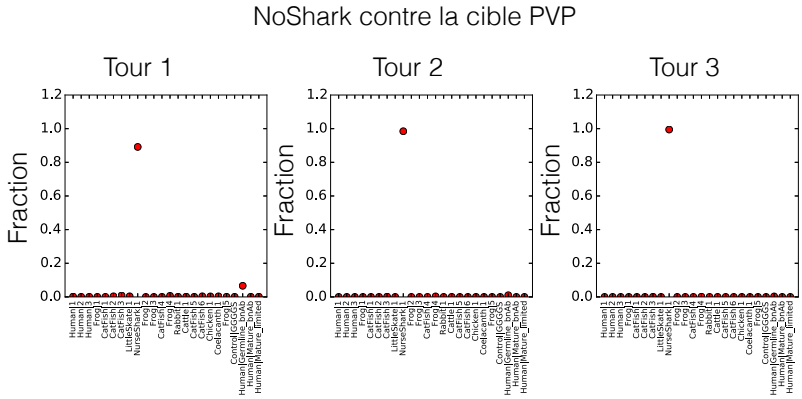


FIGURE 22: **Répartition des frameworks dans NoShark avec les tours de sélection contre PVP.** Ici aussi, dès le premier tour, NurseShark est très majoritairement sélectionné. Noté comment human germine est aussi présent au tour 1.

Frog3 contre PVP

La séquence consensus cette fois-ci est moins claire que pour NurseShark. Par contre elle est clairement différente : enrichissement pour E et V au site 1, pour S et T au site 2 et 4 et W et R au site 3 Fig. 28. Les corrélations entre sites Fig. 30, elles, sont différentes des corrélations initiales Fig.29 dès le premier tour mais évoluent quand même d'une façon cohérente avec les tours. Notez encore une fois la

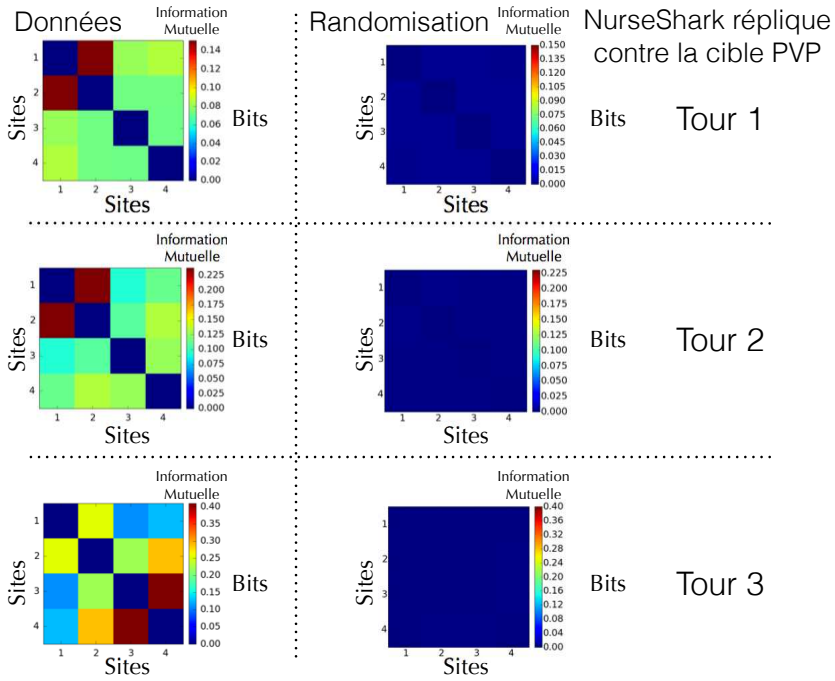


FIGURE 27: **Information mutuelle entre sites du CDR3 dans la banque Nurseshark issue de NoShark au fil des tours de sélection contre PVP.** Des corrélations significatives sont observées entre les sites. Le schéma de ces corrélations correspond à celui issu du crible de Mix21 contre PVP.

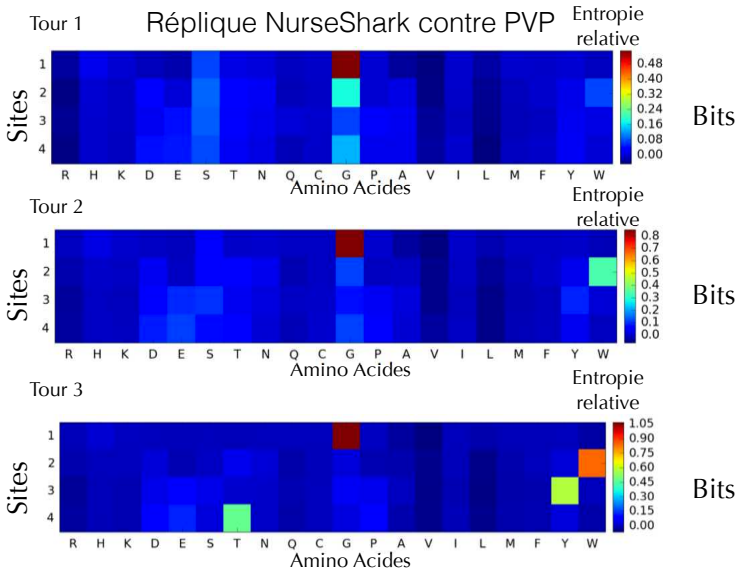


FIGURE 24: **Quantité d'information représentée par la présence d'un acide aminé particulier sur un site du CDR3 au cours du criblage de NoShark contre la cible PVP.** Une séquence consensus de type 'GWYT' est fortement sélectionnée.

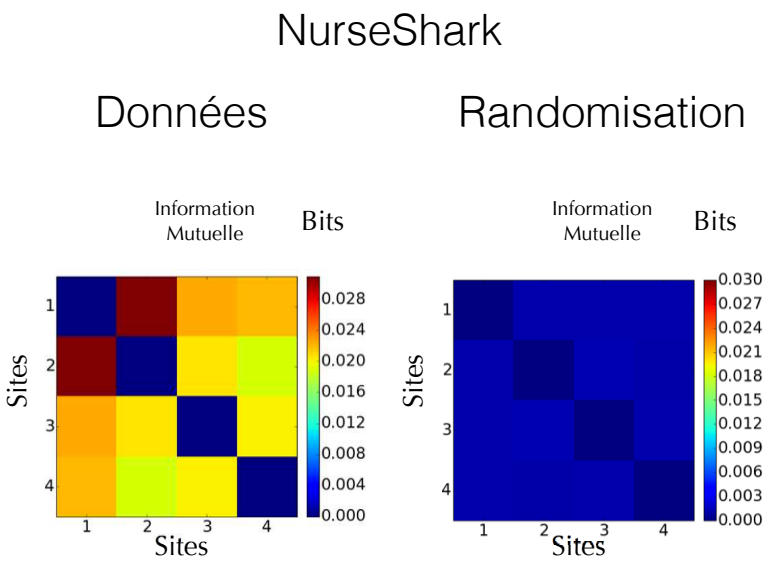


FIGURE 25: **Information mutuelle entre sites du CDR3 dans la banque initiale NurseShark.** Cette fois, les corrélations vue dans la banque initiale ne peuvent pas s'expliquer par l'échantillonnage et remettent donc en cause le caractère indépendant des sites.

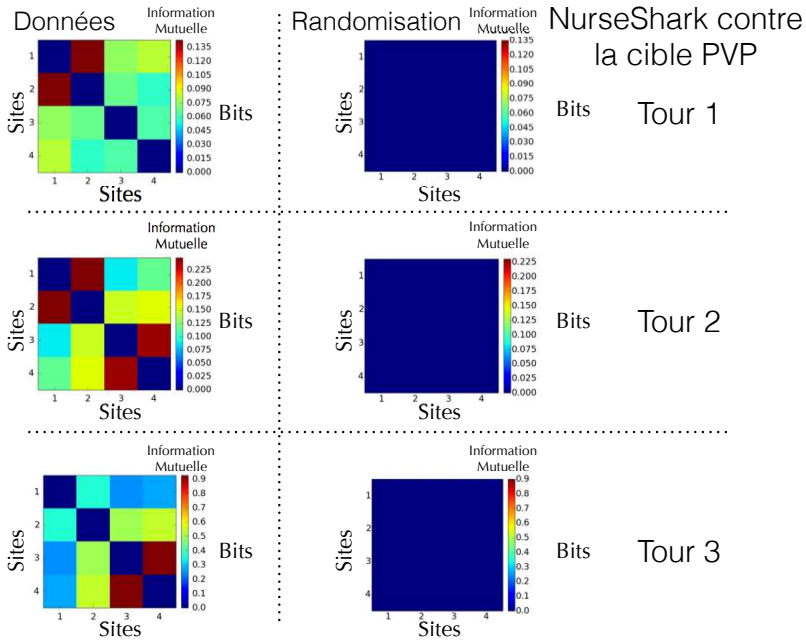


FIGURE 26: **Information mutuelle entre sites du CDR3 dans la banque Nurseshark issue de Mix21 au fil des tours de sélection contre PVP.** Des corrélations significatives sont observées entre les sites.

différence de distribution en terme de corrélation entre les deux expériences déjà présentées. Cette fois les corrélations initiales sont du même ordre de grandeurs que celles engendrées par de l'échantillonnage et ne permettent pas de réfuter l'hypothèse d'indépendance des sites. Les corrélations de Frog3 et NurseShark se ressemblent fortement Fig.29 et Fig.25.

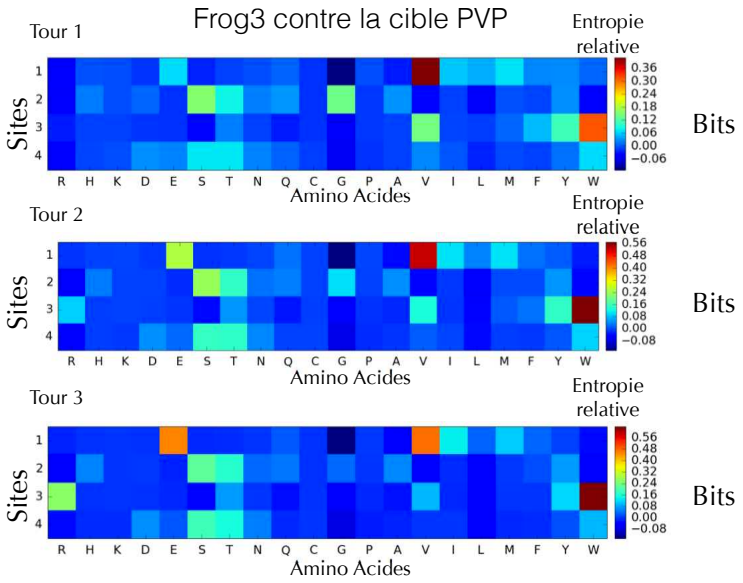


FIGURE 28: **Quantité d'information représentée par la présence d'un acide aminé particulier sur un site du CDR3 au cours du criblage de Frog3 contre PVP.**On voit effectivement un enrichissement au cours des tours pour certains acides aminés.

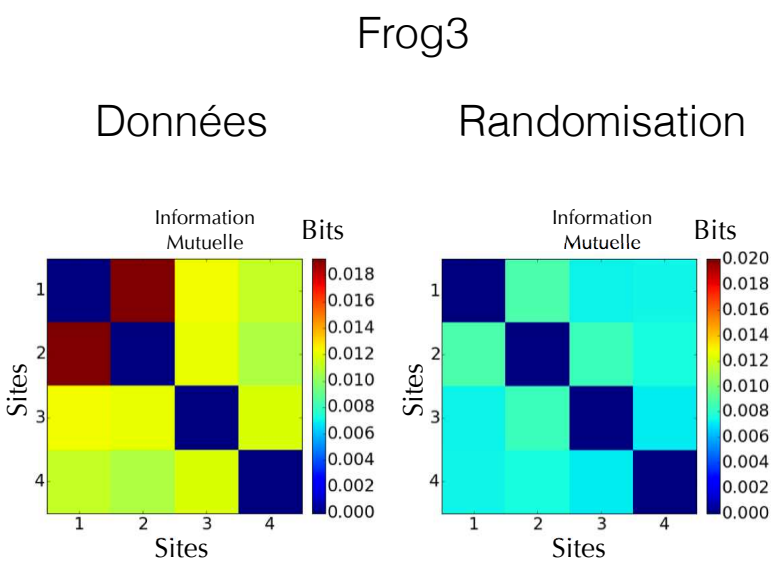


FIGURE 29: **Information mutuelle entre sites du CDR3 dans la banque initiale Frog3.** Les corrélations vues dans la banque initiale peuvent s'expliquer par l'échantillonnage.

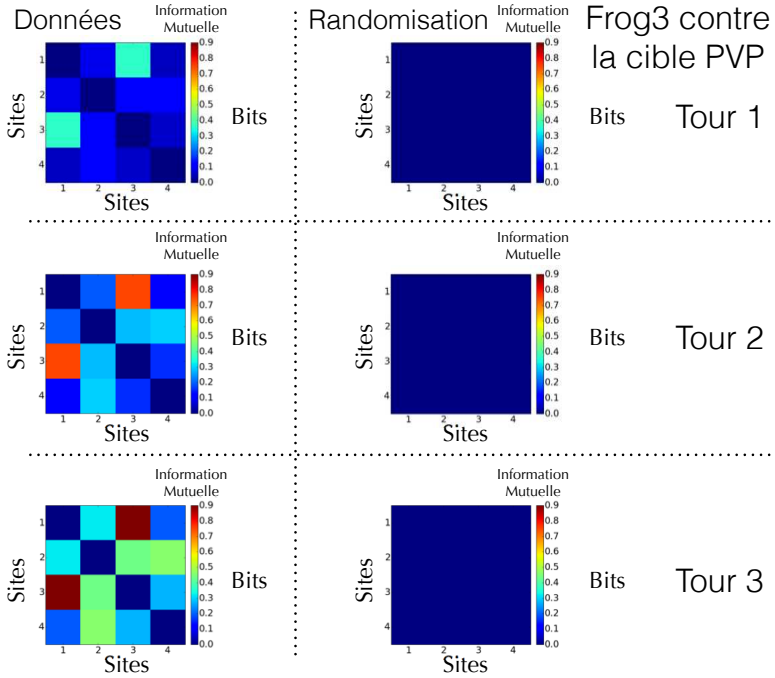


FIGURE 30: **Information mutuelle entre sites du CDR3 dans la banque Frog3 au fil des tours de sélection contre PVP.** Des corrélations significatives sont observées entre les sites.

Frog3 contre la cible noire

Il s'agit du même framework que précédemment mais sélectionné contre une autre cible. La sélection donne une séquence consensus claire : 'AELD' Fig.31. Celle-ci est foncièrement différente de la séquence consensus sélectionnée contre la cible PVP : ce qui prouve une sélection spécifique Fig.31 et Fig.28. Les corrélations entre sites sont encore une fois significatives et changent graduellement avec les tours passant d'une distribution proche de la banque initiale Fig.29, à une distribution propre à la sélection Fig.32. Même au niveau des corrélations les deux expériences sont foncièrement différentes Fig.32 et Fig.30.

Human2 contre noire

Lors de cette sélection contre noire une séquence consensus de type 'QQQQ' est sélectionnée Fig.33. Des corrélations significatives sont observées Fig.35. La banque initiale présente des corrélations mais non significatives devant celles du à l'échantillonnage Fig.34 : pour autant la distribution des corrélations est très similaire à Fig.29 et Fig.25.

NoFramework contre noire

Cette banque a été initialement sous échantillonnée : il est impossible de faire une analyse en terme d'entropie relative. On a donc fait une analyse en terme d'entropie : celle-ci ne montre aucune sélection visible Fig.36. Des corrélations significatives sont visibles Fig.37. Mon avis sur cette expérience est qu'elle n'a pas bien fonctionné, ou

Entropie : $D(X_k) = \sum_{i \in aa} -P(x_{i,k}) \log_2(P(x_{i,k})$

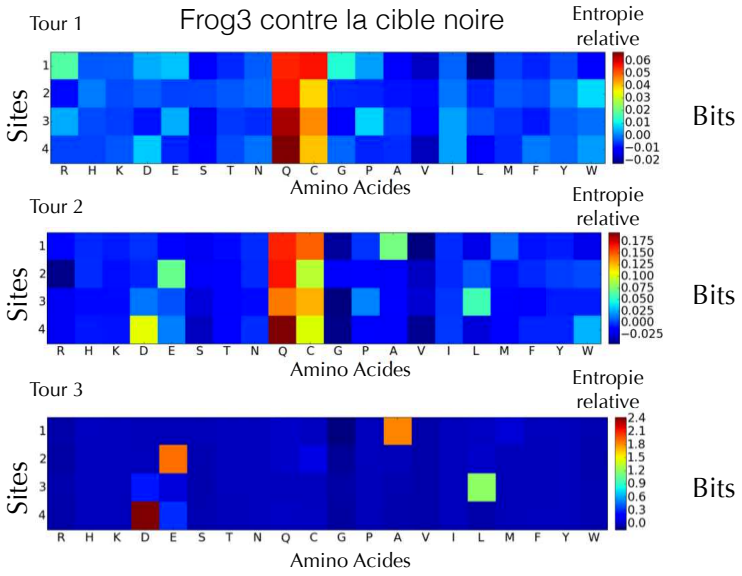


FIGURE 31: **Quantité d'information représentée par la présence d'un acide aminé particulier sur un site du CDR3 au cours du criblage de Frog3 contre la cible noire.**On voit effectivement un enrichissement au cours des tours pour certains acides aminés.

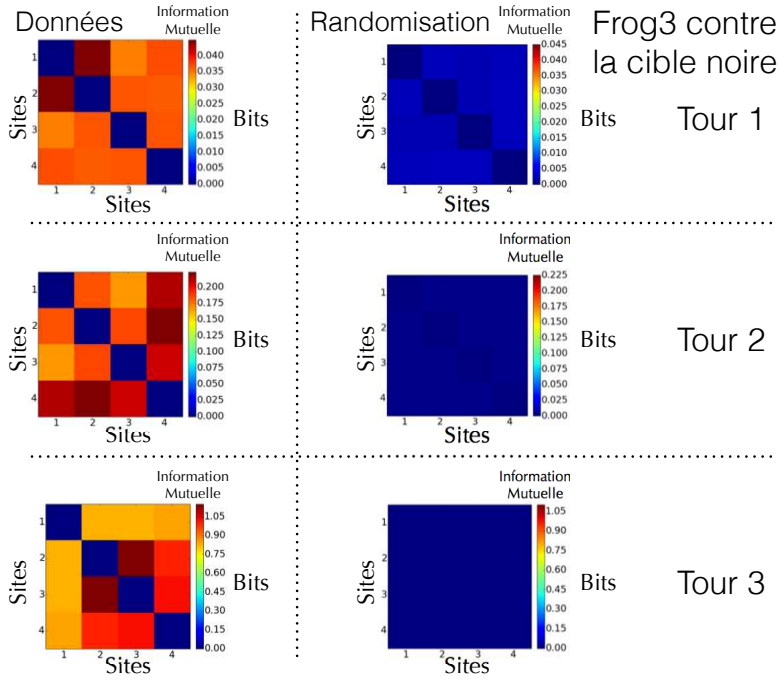


FIGURE 32: **Information mutuelle entre sites du CDR3 dans la banque Frog3 au fil des tours de sélection contre la cible noire.** Des corrélations significatives sont observées entre les sites.

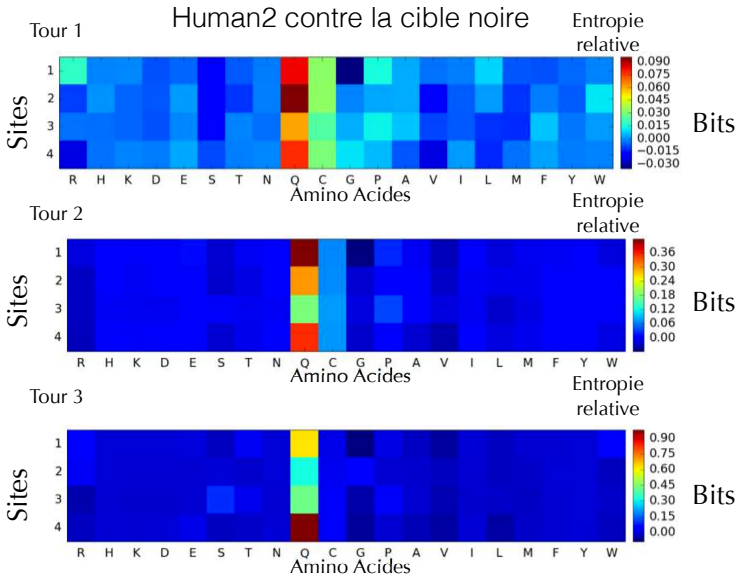


FIGURE 33: **Quantité d'information représentée par la présence d'un acide aminé particulier sur un site du CDR3.** On remarque que la quantité d'information augmente avec le nombre de tour et converge, d'une façon continue, vers une séquence enrichie en glutamine Q.

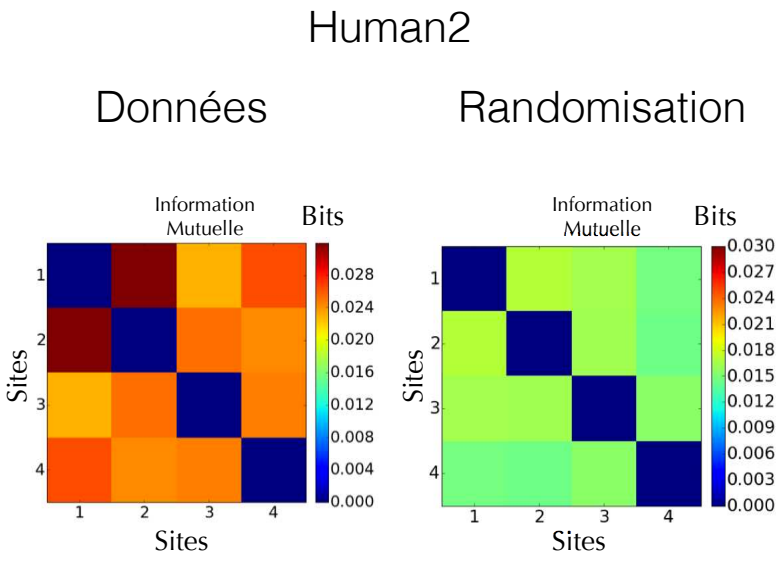


FIGURE 34: **Information mutuelle entre sites du CDR3 dans la banque initiale human2.** On remarque que l'information mutuelle entre les données et les données simulées sur la base du modèle sans corrélations mais avec échantillonnage sont du même ordre de grandeur : les corrélations vues dans cette banque initiale peuvent s'expliquer par l'échantillonnage.

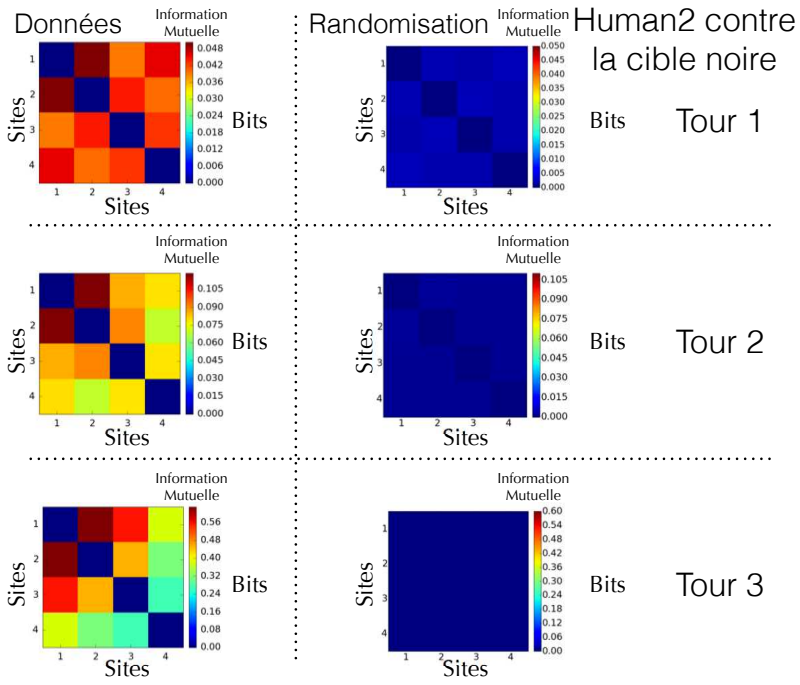


FIGURE 35: **Information mutuelle entre sites du CDR3 dans la banque human2 au fil des tours de sélection.**Des corélations significatives sont observées entre les différents sites du CDR3.

alors très faiblement : les données montrant la sélectivité globale et totale : à chaque tour (données non montrées ici), le prouve ainsi que les données section *Robustesse de l'analyse*. Le fait que des corrélations existent peut être expliqué si on suppose que sont là les corrélations présentes initialement dans une banque qui comme le montre Fig.19 ou Fig.18 est initialement problématique. Pour autant, elle fera quand même partie du reste de l'analyse et de la robustesse pour jouer le rôle d'un contrôle et remettre en perspective les autres résultats.

Sélectivité globale : Ratio entre le nombre phages présents avant et après sélection.

Mix21bis contre noire.

Dans cette expérience, encore une fois, un framework domine : il s'agit de Chicken Fig.38. Pour autant, comme d'autres frameworks sont présents aux différents tours, et cela d'une manière non négligeable, nous ne nous sommes pas intéressés aux corrélations entre sites et à l'enrichissement de certains acides aminés à certaines positions, car ceux-ci sont framework dépendants (les résultats par framework sont disponibles en annexe).

Mix24 contre noire.

Dans cette expérience, encore une fois, un framework domine : il s'agit de Human Germline Fig.39. Pour les mêmes raisons que Mix21bis contre noire, l'enrichissement en acides aminés particuliers et en corrélations entre sites n'est pas présenté. Encore une fois une analyse similaire par framework est, lorsque c'est possible, dispo-

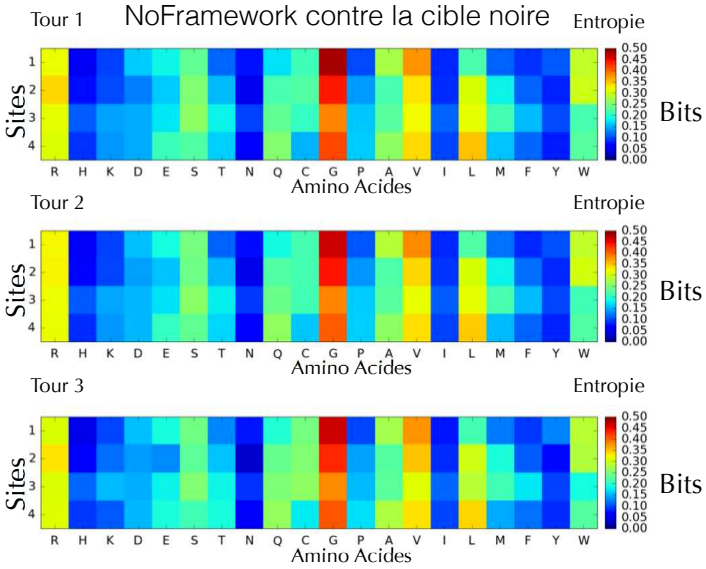


FIGURE 36: **Quantité d'information représentée par la présence d'un acide aminé particulier sur un site du CDR3.**On ne voit pas de changement significatif dans la distribution des acides aminés par site.

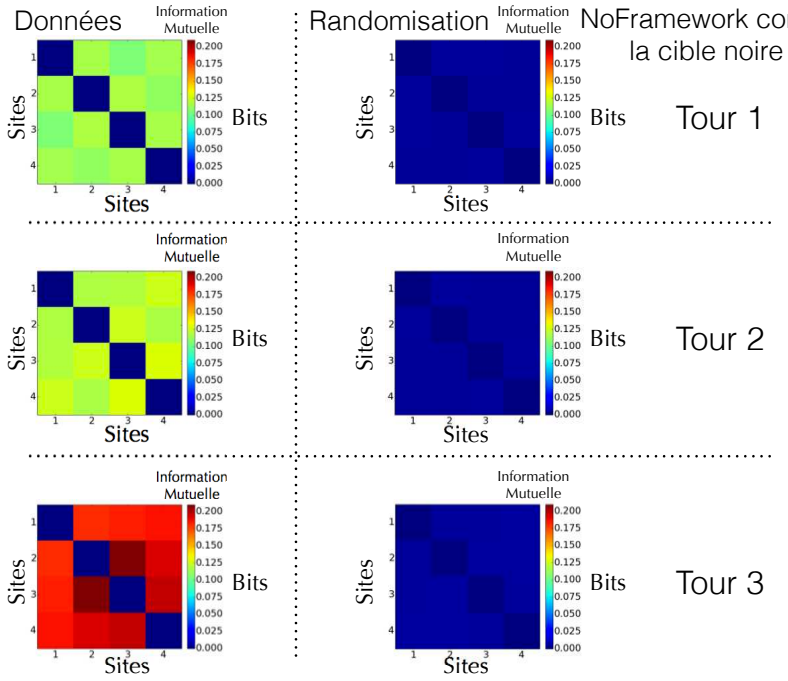


FIGURE 37: **Information mutuelle entre sites du CDR3 dans la banque NoFramework au fil des tours de sélection.** Des corrélations significatives sont observées entre les sites.

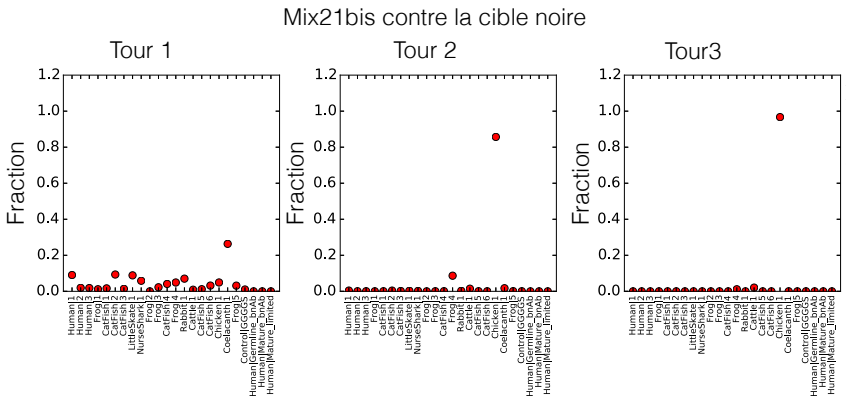


FIGURE 38: **Répartition des frameworks dans Mix 21bis avec les tours de sélection contre noire.** Au fur et à mesure des tours le framework chicken l'emporte. Pourtant parmi les séquences qui ont une forte sélectivité subsiste certains fragments portant le framework : Chicken, Cattle, Human1, NurseShark , Littleskate, NoFramework

nible en annexe.

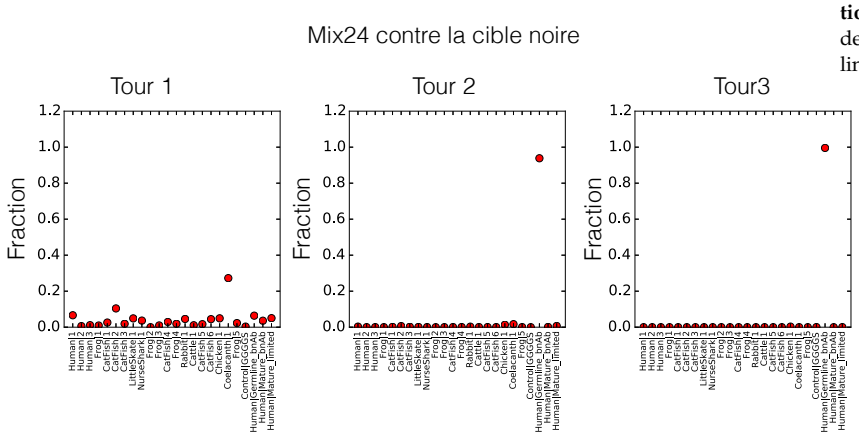


FIGURE 39: **Répartition des frameworks dans Mix 24 avec les tours de sélection contre noire.** Au fur et à mesure des tours le framework Human Germline l'emporte.

Mix24 contre bleu et contre rouge.

Ces deux sélections ont été mal séquencées : seulement 50 000 CDR3 ont pu être retrouvées et aucune analyse plus poussée n'a été entreprise Fig.40 et Fig.41.

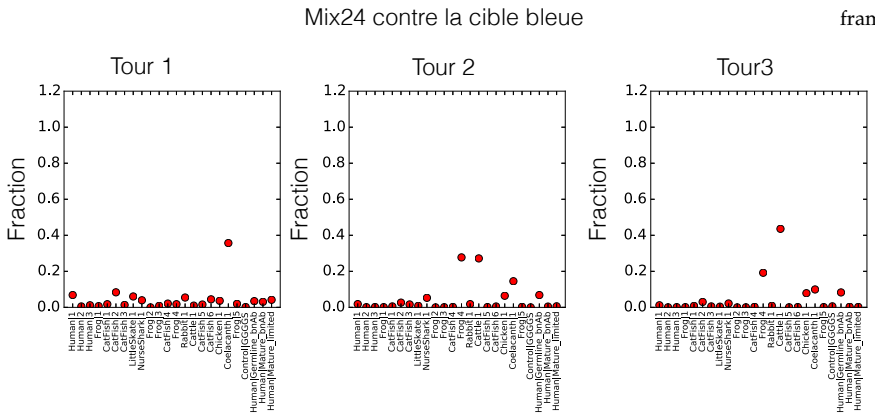


FIGURE 40: **Répartition des frameworks dans Mix 24 avec les tours de sélection contre bleue.** Il n'y a pas vraiment de framework qui l'emporte.

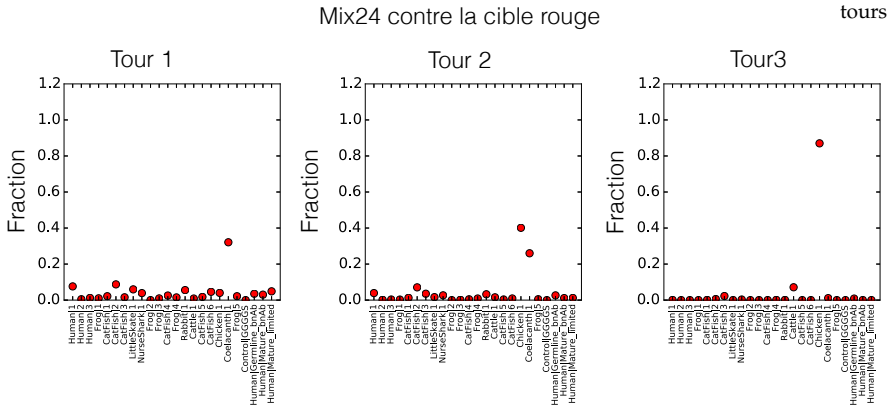


FIGURE 41: **Répartition des frameworks dans Mix 24 avec les tours de sélection contre rouge.** Au fur et à mesure des tours le framework Chicken l'emporte.

Analyse en valeurs extrêmes

Dans cette partie on calcule les sélectivités des séquences dans les différentes expériences présentées ci-avant Eq.9 et on procéde à une analyse en valeurs extrêmes comme présenté dans le chapitre *Réponse d'un système à la sélection* section *Analyse en valeurs extrêmes*. Le calcul des sélectivités se fait entre les tours 2 et 3. Pour cela deux contraintes sont ajoutées aux jeux de données qui avaient déjà étét triés pour la qualité de lecture :

- On ne prend en compte dans l'analyse que les séquences présentent à plus de 10 exemplaires au tour 2 (comme ensuite on ne prendra que les meilleurs sélectivités on peut faire l'hypothèse que cette condition implique plus de 10 exemplaires au tour 3). Cette condition a pour but de contraindre la qualité de l'estimation des sélectivités.
- On ne fait l'analyse que si suffisamment de points répondent à la condition précédente : on ajuste le dernier $\hat{\kappa}$ sur au moins 50 points et il faut au moins 100 autres estimations de $\hat{\kappa}$ pour pouvoir déceler un plateau. En tout il faut donc qu'au moins 150 sélectivités répondent au critère 1.

MIx 21 et NoShark contre PVP.

L'analyse en valeurs extrêmes est aussi plutôt reproductible. La section *Robustesse de l'analyse* en montrera les limites. Puisque Mix21 PVP a été plus séquencé que NoShark PVP, on peut s'assurer que l'incertitude sur les sélectivités calculés ne joue pas un rôle trop important sur l'analyse, en ré échantillonnant Mix21 PVP au même taux que NoShark PVP. L'analyse n'est pas affectée : les graphes se ressemblent fortement (linéarité dans le log log plot et plateau de l'estimateur) et semblent indiquer un $\hat{\kappa}$ aux alentours de 0.5 Fig.42 et Fig.43.
Le crible de NoShark contre PVP donne un ajustement : à partir de $\hat{s}^* = 1.6 * 10^{-3}$ tel que $\hat{\kappa} = 0.28 \pm 0.14$, ce qui équivaut à un $\hat{\tau} = 2.8 * 10^{-4} \pm 0.5 * 10^{-4}$. Le crible de Mix21 donne un ajustement : à partir de $\hat{s}^* = 1.8 * 10^{-3}$ tel que $\hat{\kappa} = 0.42 \pm 0.22$. Dans les deux cas la P-value associée à une erreur de différenciation entre Gumbel et Fréchet est de moins de 10^{-4}. Comme on a plus de séquences pour Mix21 contre PVP, on peut se permettre de regarder à quel point ce seuil de plus de 10 au tour 2 est important : si on passe à 50 l'ajustement permet d'avoir un $\hat{\kappa} = 0.35 \pm 0.33$ et une P-value de $6 * 10^{-4}$.

Frog3 contre PVP

Lorsque l'on fait l'analyse en terme de valeurs extrêmes pour Frog3 contre PVP Fig.45, le caractère linéaire du graphe log-log sélectivité vs rang, qui est un premier indice dans la détermination de la classe d'universalité, n'est pas très clair. L'ajustement donne un

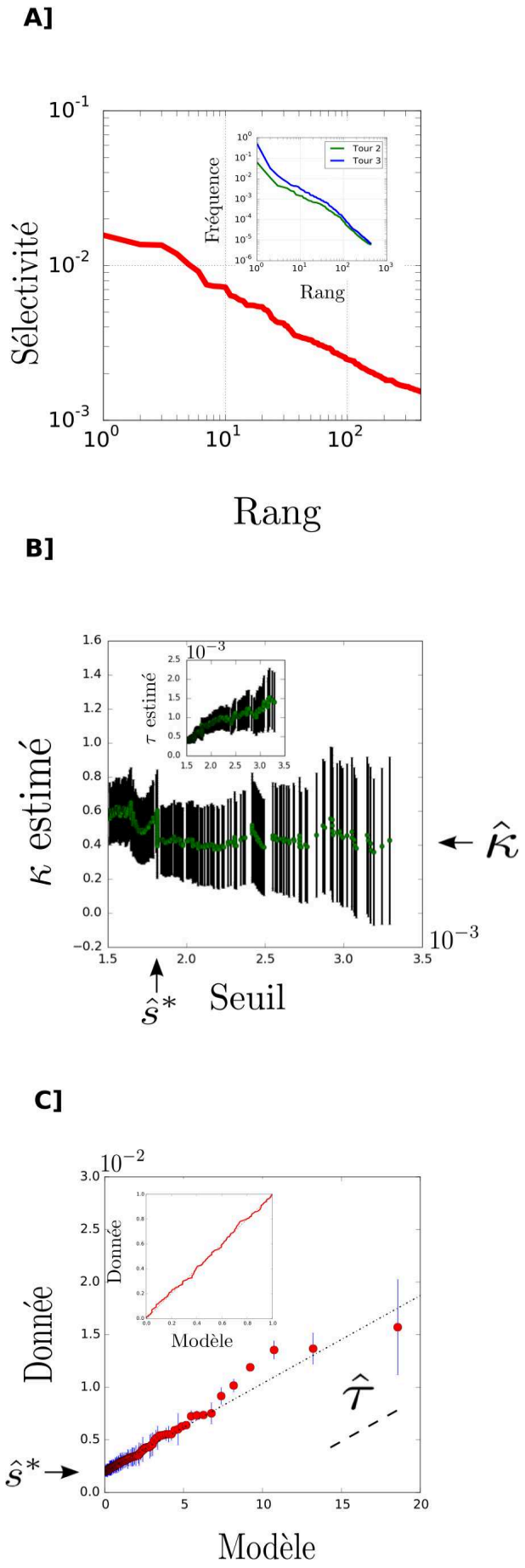


FIGURE 42: **Représentation de la distribution de sélectivités, estimation des différents paramètres de la distribution de sélectivitsé et ajustement du modèle sur les données pour Mix21 contre PVP. A]** Rang vs Sélectivité. On remarque que dans ce graphe en log-log la sélectivité décroit linéairement avec le rang. Encart : Fréquences vs rang au tour 2 et au tour 3. Il est à noter qu'ici le rang est tour dépendant : une séquence n'a pas forcément le même rang au tour 2 et 3.**B]** Dépendance des différents estimateurs avec le choix du seuil. On choisit le seuil de sorte que l'estimateur pour κ soit constant dans les barres d'erreurs. Comme expliqué précédemment les barres d'erreurs sont les intervalles à 95 pour cent de confidence sur l'estimateur, dans l'hypothèse où l'estimateur est gaussien. On prend alors le τ associé à ce seuil en s'assurant qu'a partir du seuil choisit τ suivent bien κ. Dans ce cas précis : $\hat{s}^* = 1.8 * 10^{-3}$ équivaut à un $\kappa = 0.42$ stable, à partir de ce point on peut estimer une pente de τ_{estime} vs seuil = 0.7/1.5 = 0.46. **C]** Comparaison entre le modèle issue de l'ajustement et les données. Figure principale : les sélectivités prévues par le modèle et les sélectivités issues de l'expérience sont comparés. La plupart des points issus du modèle et données sont alignés sur une droite de pente τ et d'origine le seuil choisi, signe que l'ajustement est acceptable. Dans ce cas le seul paramètre donné au modéle est κ, le recentrage et le rescaling se faisant par l'intermédiare de l'ordonnée à l'origine et de la pente, par définition. Les barres d'erreurs sont ici une estimation de la variabilité des s due à l'échantillonnage. Encart : cette fois ce sont les distributions cumulatives qui sont comparées. Dans ce cas les 3 paramétres sont fixés par le modéle. Encore une fois le fait que les points forment une ligne droite de pente 1 et d'origine 0 est gage d'un ajustement acceptable.

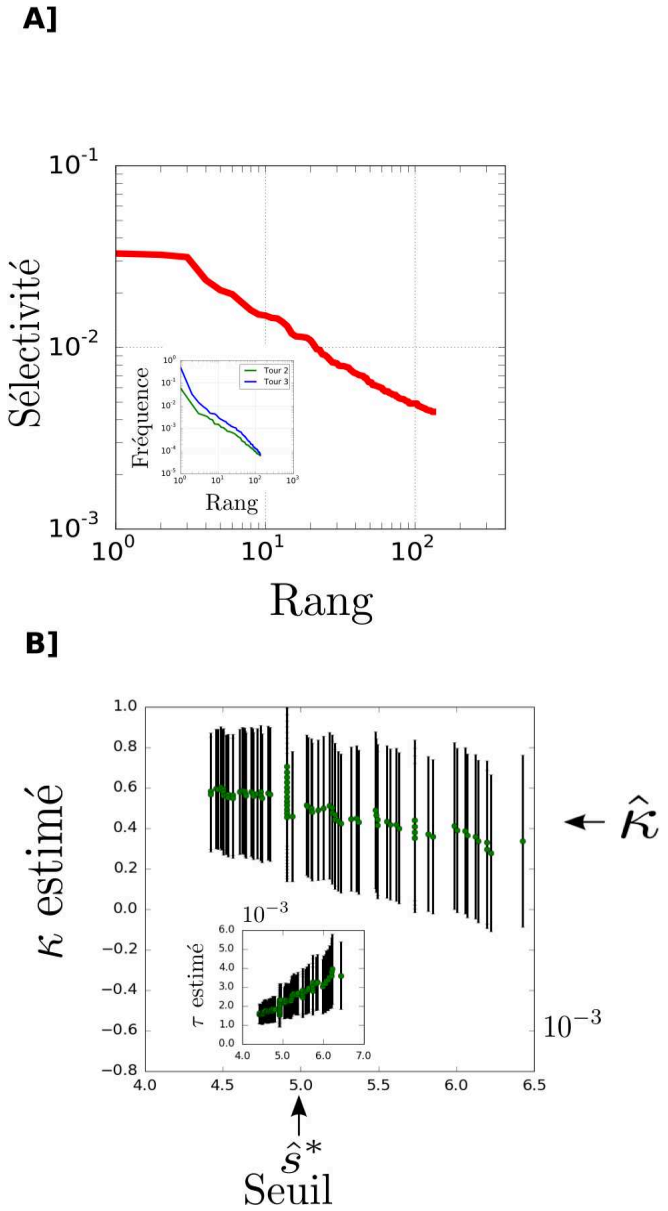


FIGURE 43: **Représentation de la distribution de sélectivités et estimation des différents paramètres de la distribution de sélectivités pour le criblage de Mix21 contre PVP, en réduisant artificiellement le taux de lecture pour être comparable aux autres expériences. A]** Rang vs Sélectivité. Encart : Fréquences vs rang au tour 2 et au tour 3. **B]** Dépendance des différents estimateurs avec le choix du seuil. On ne peut pas vraiment définir de plateau pour l'estimateur $\hat{\kappa}$. Ces deux graphes sont très proches de ceux Fig.42, et montrent que la qualité de l'estimation des fréquences ne jouent pas trop sur l'analyse.

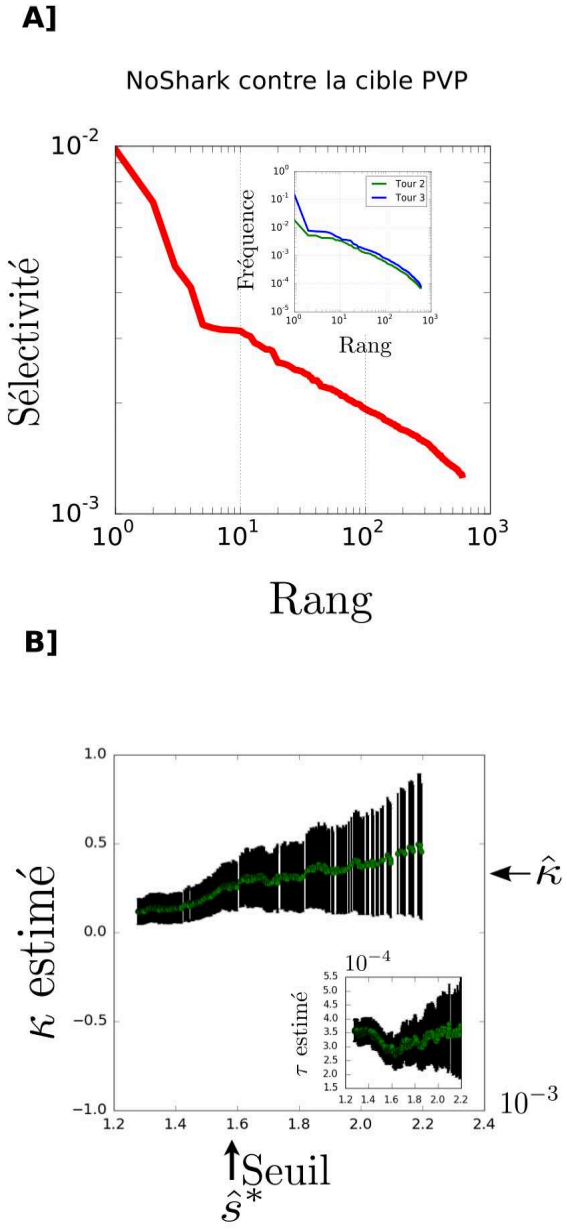


FIGURE 44: **Représentation de la distribution de sélectivités, estimation des différents paramètres de la distribution de sélectivités et ajustement du modèle sur les données pour le criblage de NoShark contre PVP. A]** Rang vs Sélectivité. On remarque que dans ce graphe en log-log la sélectivité décroit linéairement avec le rang. Encart : Fréquences vs rang au tour 2 et au tour 3. **B]** Dépendance des différents estimateurs avec le choix du seuil. On peut grossièrement définir un plateau à partir de $\hat{s}^* = 1.6 * 10^{-3}$ tel que $\hat{\kappa} = 0.28 \pm 0.14$. Ceci équivaut à un $\hat{\tau} = 2.8 * 10^{-4} \pm 0.5 * 10^{-4}$. **C]** Comparaison entre le modèle issue de l'ajustement et les données. Figure principale : l'ajustement en terme de sélectivité est acceptable. Encart : l'ajustement en terme distribution cumulative est lui aussi acceptable. Les simulations montrent qu'il y a moins d'une chance sur dix milles, étant donné nos données, de ne pas pouvoir différencier cette distribution de type Fréchet avec le modèle nul de type Gumbel.

$\hat{\kappa} = 0.07 \pm 0.1$ très proche de 0. Lorsque l'on essaie de faire la différence entre cette distribution de type Weibull et son homologue exponentielle (Gumbel), on se rend compte qu'il y 97% de chances de se tromper. On considérera donc cette distribution comme étant de type Gumbel avec un $\hat{\tau}_{exp} = 4.7 * 10^{-4} \pm 0.3 * 10^{-4}$. L'ajustement de la cumulative et de la fonction quantile est acceptable : seul le cas $\hat{\kappa} = 0.07 \pm 0.1$ est présenté mais les distributions sont tellement proches que l'ajustement est identique dans l'hypothèse exponentielle.

Frog3 contre noire

Il est impossible de procéder à l'analyse en valeurs extrême compte tenu du nombre trop petit de points restant après l'étape où l'on ne prend que les séquences présentes au moins 10 fois au tour 2.

Human2 contre noire

On procède alors à l'analyse en terme de valeurs extrêmes Fig.46. On s'aperçoit que pour un seuil supérieur à $\hat{s}^* = 1.3 * 10^{-3}$ le maximum likelihood estimator pour $\hat{\kappa}$ se stabilise à 0.5 ± 0.2, ce qui correspond à un $\hat{\tau} = 2.7 * 10^{-4} \pm 0.7 * 10^{-5}$.Pour ces paramètres l'ajustement est acceptable aussi bien pour la distribution cumulative que pour la distribution des quantiles.

Enfin, des simulations montrent qu'il y a moins d'une chance sur dix milles, étant donné nos données, de ne pas pouvoir différencier cette distribution de type Fréchet avec le modèle nul de type Gumbel.

NoFramework contre noire

Lorsque l'on fait l'analyse en terme de valeurs extrêmes Fig.47, on se rend compte cette fois-ci que l'on ne retrouve plus du tout le caractère linéaire du graphe log-log sélectivité vs rang. L'ajustement donne un $\hat{\kappa} = 0.03 \pm 0.08$ très proche de 0. Lorsque l'on essaie de faire la différence entre cette distribution de type Fréchet et son homologue exponentielle (Gumbel), on se rend compte qu'il y 73% de chances de se tromper. On considérera donc cette distribution comme étant de type Gumbel avec un $\hat{\tau}_{exp} = 10^{-4} \pm 0.5 * 10^{-4}$.

Mix21bis contre noire

L'analyse en terme de valeurs extrêmes montre des résultats très moyens en terme d'ajustement pour le modèle Weibull trouvé par l'estimateur (à partir de $\hat{s}^* = 3.2 * 10^{-3}$ tel que $\hat{\kappa} = -0.38 \pm 0.21$.) Fig.48. L'ajustement n'est d'ailleurs pas bon non plus pour le modèle Gumbel : $\hat{\tau}_{exp} = 8.710^{-3} \pm 2 * 10^{-3}$ Fig.49 (cette fois les deux ajustements sont présentés à la différence de la sous-section *Forg3 contre PVP* et *No Framework contre noire* car $\hat{\kappa} = -0.38 \pm 0.21$ est foncièrement différent de 0).

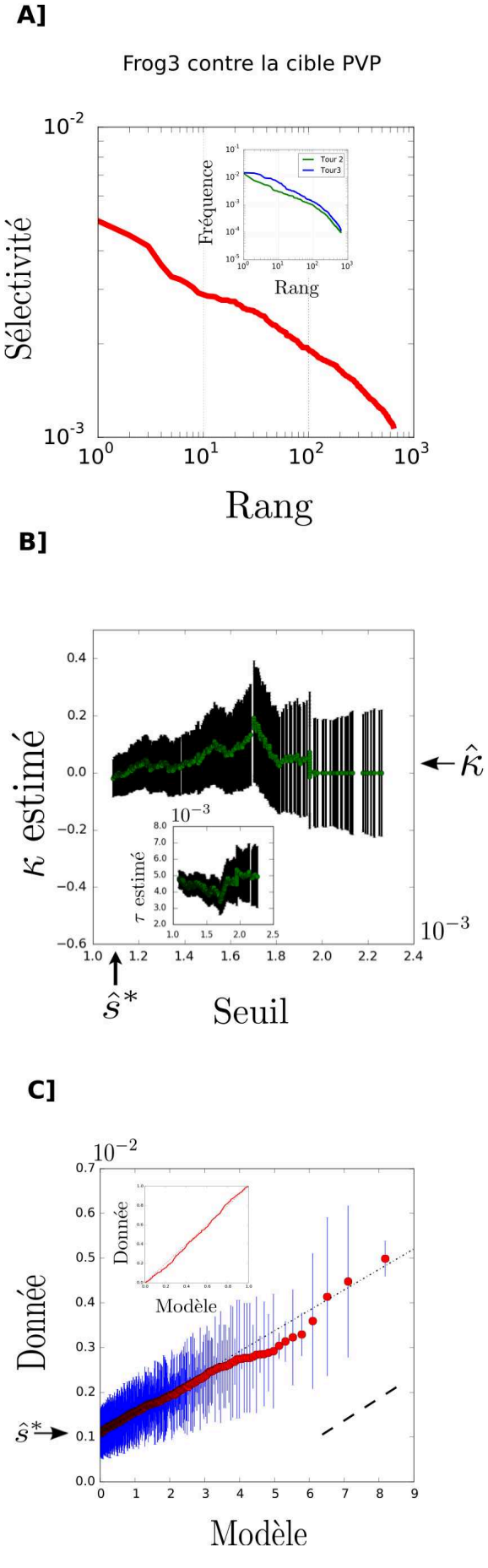


FIGURE 45: **Représentation de la distribution de sélectivités, estimation des différents paramètres de la distribution de sélectivités et ajustement du modèle sur les données pour le criblage de Frog3 contre PVP. A]** Rang vs Sélectivité. On remarque que dans ce graphe en log-log la sélectivité décroit pseudo linéairement avec le rang. Encart : Fréquences vs rang au tour 2 et au tour 3. **B]** Dépendance des différents estimateurs avec le choix du seuil. On peut définir un plateau à partir de $\hat{s}^* = 1.1 * 10^{-3}$ tel que $\hat{\kappa} = -0.07 \pm 0.1$. Ceci équivaut à un $\hat{\tau} = 4.5 * 10^{-4} \pm 0.5 * 10^{-4}$**C]**Comparaison entre le modèle issu de l'ajustement et les données. Figure principale : l'ajustement en terme de sélectivité est acceptable. Encart : l'ajustement en terme distribution cumulative est lui aussi acceptable. Par contre, comme on peut le voir l'ajustement est très proche d'une exponentielle car $\hat{\kappa}$ est très proche de 0. La probabilité de ne pas pouvoir cette distribution avec une distribution exponentielle de type Gumbel avec $\hat{\tau}_{exp} = 4.7 * 10^{-4} \pm 0.3 * 10^{-4}$ est de 97 %.

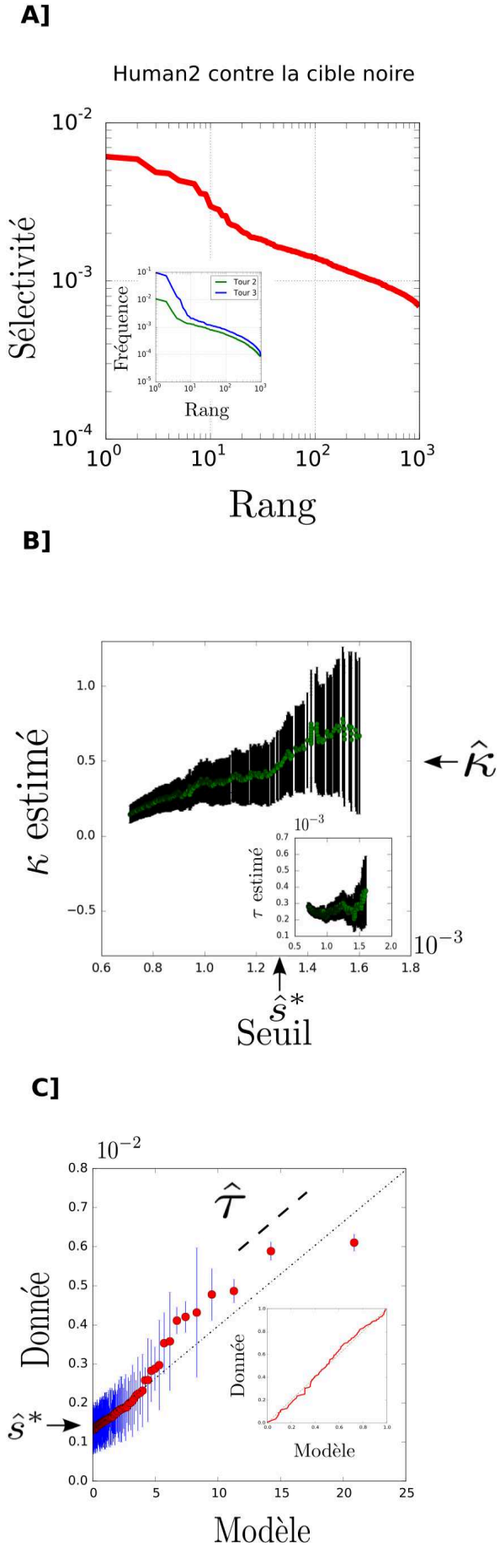


FIGURE 46: **Représentation de la distribution de sélectivités, estimation des différents paramètres de la distribution de sélectivités et ajustement du modèle sur les données pour le criblage de Human2 contre noire. A]** Rang vs Sélectivité. On remarque que dans ce graphe en log-log la sélectivité décroit linéairement avec le rang. Encart : Fréquences vs rang au tour 2 et au tour 3. **B]** Dépendance des différents estimateurs avec le choix du seuil. On peut définir un plateau à partir de $\hat{s}^* = 1.3 * 10^{-3}$ tel que $\hat{\kappa} = 0.5 \pm 0.2$. **C]**Comparaison entre le modèle issue de l'ajustement et les données. Figure principale : l'ajustement en terme de sélectivités est acceptable. Encart : l'ajustement en terme distribution cumulative est lui aussi acceptable. Les simulations montrent qu'il y a moins d'une chance sur dix milles, étant donné nos données, de ne pas pouvoir différencier cette distribution de type Fréchet avec le modèle nul de type Gumbel.

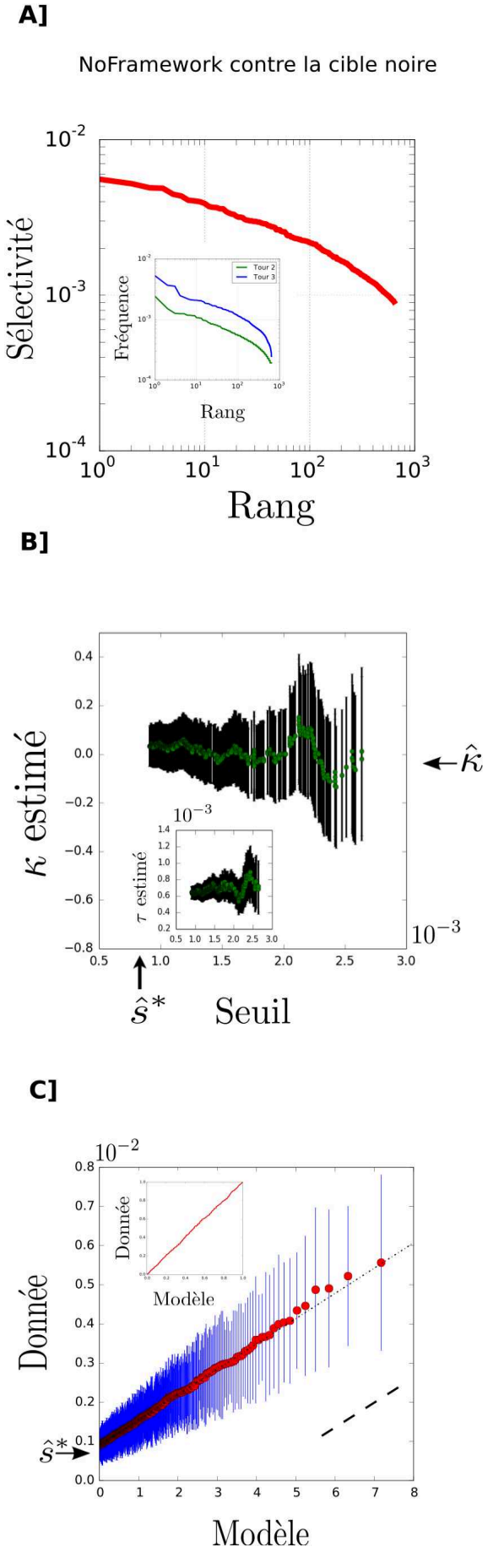


FIGURE 47: **Représentation de la distribution de sélectivités, estimation des différents paramètres de la distribution de sélectivités et ajustement du modèle sur les données pour NoFramework contre noire .A]** Rang vs Sélectivité. Cette fois on ne retrouve pas de relation linéaire entre le rang et la sélectivité en log-log Encart : Fréquences vs rang au tour 2 et au tour 3.**B]** Dépendance des différents estimateurs avec le choix du seuil. L'estimateur semble atteindre un plateau depuis le début : on s'est assuré de ne prendre que les séquences qui augmentaient leur part de la population entre les tours 2 et 3. L'ajustement donne à partir de $\hat{s}^* = 0.8 * 10^{-3}$ un $\hat{\kappa} = 0.03 \pm 0.08$. Ceci équivaut à un $\hat{\tau} = 6.4 * 10^{-4} \pm 0.7 * 10^{-4}$**C]**Comparaison entre le modèle issue de l'ajustement et les données. Figure principale : les sélectivités prévue par le modèle et les sélectivités issues de l'expérience sont comparés et montrent un très bon accord. Encart : cette fois se sont les distributions cumulatives qui sont comparées : de la même façon que pour la comparaison des quantiles, on peut dire que l'ajustement est acceptable. Par contre, comme on peut le voir l'ajustement est très proche d'une exponentielle car $\hat{\kappa}$ est très proche de 0. La probabilité de confondre cette distribution avec une distribution exponentielle de type Gumbel avec $\hat{\tau}_{exp} = 10^{-4} \pm 0.5 * 10^{-4}$ est de 73 pour-cent .

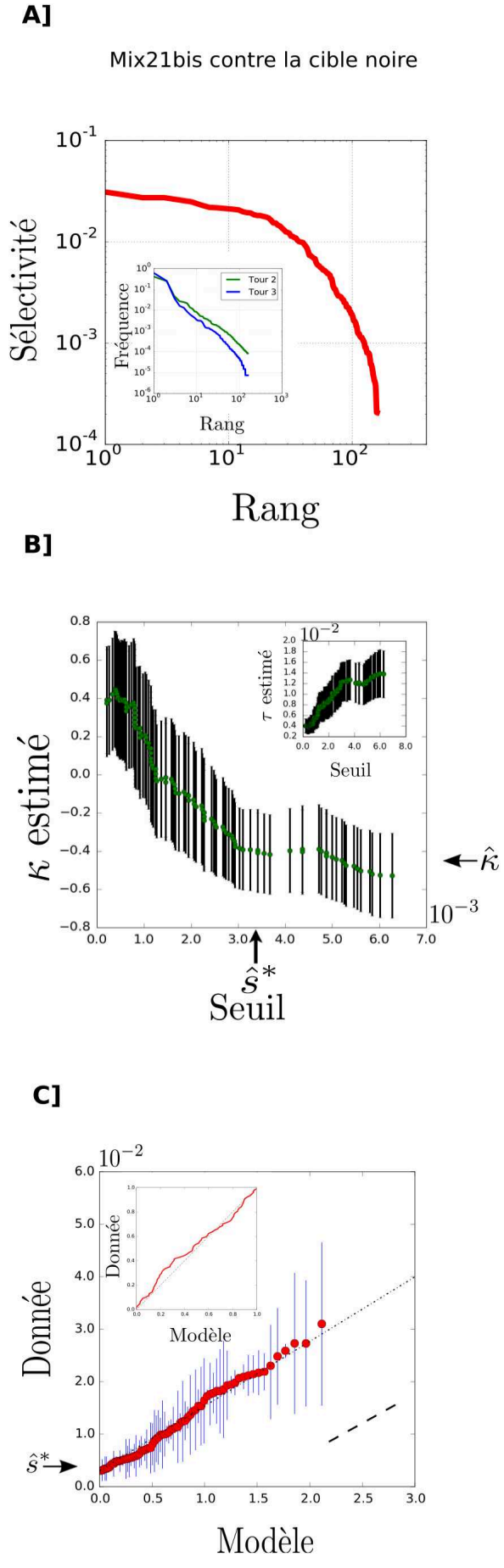


FIGURE 48: **Représentation de la distribution de sélectivités, estimation des différents paramètres de la distribution de sélectivités et ajustement du modèle sur les données pour le criblage de Mix2bis contre noire.A]** Rang vs Sélectivité. On remarque que dans ce graphe en log-log la sélectivité est plate sur 1 ordre de grandeurs du rang puis tombe rapidement. Encart : Fréquences vs rang au tour 2 et au tour 3. **B]** Dépendance des différents estimateurs avec le choix du seuil. On peut grossièrement définir un plateau à partir de $\hat{s}^* = 3.2 * 10^{-3}$ tel que $\hat{\kappa} = -0.38 \pm 0.21$. Ceci équivaut à un $\hat{\tau} = 1.2 * 10^{-2} \pm 0.4 * 10^{-2}$. Par contre la relation entre le seuil et τ n'étant absolument pas respecté, on peut s'attendre à ce que ces données ne soient pas décrites par une GPD. **C]**Comparaison entre le modèle issue de l'ajustement et les données. Figure principale : l'ajustement en terme de sélectivités est passable. Encart : l'ajustement en terme distribution cumulatives est mauvais. Les simulations montrent qu'il y a 5% de chances, étant donné nos données, de ne pas pouvoir différencier cette distribution de type Fréchet avec le modèle nul de type Gumbel.

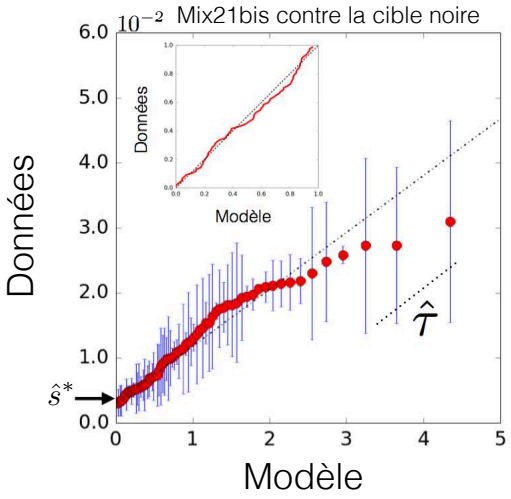


FIGURE 49: **Ajustement exponentiel du modèle sur les données pour le criblage de Mix2bis contre noire.** L'ajustement de la cumulative et de la fonction quantile sont mauvais pour cette expérience.

Mix24 contre noire, rouge et bleu

Il n'y a pas assez de données pour faire l'analyse en valeurs extrêmes.

Robustesse de l'analyse

L'analyse décrite ci- avant repose sur un certain nombre de règles, notamment concernant la définition d'une séquence bien lue et d'un coefficient de sélectivité bien estimé. Il est primordial de regarder à quel point les conclusions que l'on tire de cette analyse peuvent dépendre du jeu de paramètres utilisés pour définir le jeu de données à analyser. Nous allons vérifier la :

- Robustesse vis à vis de l'erreur acceptable pour le calcul d'une sélectivité.
- Robustesse vis à vis de la qualité de lecture des séquences.
- Robustesse vis à vis des rounds de sélection utilisés pour le calcul des sélectivités.

Les deux premiers points sont traités ensemble.
Dans toute l'analyse présentée ci-avant, seules les séquences présentes plus de 10 fois au tour 2 sont gardées pour l'analyse en valeurs extrêmes. On pourrait aussi déterminer un autre critère qui serait, lui, directement lié à la confiance que l'on accorde à la sélectivité calculée. On a alors le choix entre mettre un seuil sur l'erreur relative ou sur l'erreur absolue. Dans les deux cas on n'impose pas les mêmes contraintes : l'erreur relative va être plus permissive sur les grandes sélectivités et ce sera le contraire pour l'erreur absolue. On décide donc de refaire l'analyse avec des jeux de données différents de ceux présentés ci-avant : on ne garde que les sélectivités dont on sait l'erreur relative inférieure à 30%, ou les sélectivités dont l'erreur absolue est plus petite que l'erreur expérimentale. Définir l'erreur

expérimentale est un point difficile, pour autant on peut l'estimer globalement. En effet dans nos différents tours de sélection, certaines séquences présentent un codon stop. L'anticorps étant en amont de la protéine de surface PIII, trouver un codon stop dans nos données peut vouloir dire 2 choses : soit il s'agit là de bruit de sélection (des anticorps non sélectionnés se retrouvent par chance présents après sélection), soit il s'agit d'un bruit de lecture (il ne s'agit pas réellement de codon stop). Il devrait être possible de faire la différence entre ces deux types de bruits : dans le cas du bruit sélection la fréquence des séquences ayant un codon stop devrait diminuer avec les tours de sélection, dans le cas du bruit de lecture il devrait être plus ou moins constant (plus ou moins...). Malheureusement, on ne voit rien d'aussi flagrant : la fréquence des séquences ayant un codon stop tend bien à diminuer avec les tours, mais le fait que toute les séquences avec un codon stop dans Mix21 contre PVP et NoShark contre PVP soit des séquences NurseShark fait plutôt penser à des erreurs de lectures. Même si on ne sait pas de quel type de bruit il s'agit, on a tout de même calculé sa sélectivité (s_{noise}) pour s'en servir comme borne pour l'erreur absolue des autres sélectivités : les s que l'on prend sont très supérieur à s_{noise} et contraindre Δs revient à s'assurer que les s que l'on regarde soient significativement au dessus du bruit.

Deux tableaux récapitulent les choix de seuil sur les sélectivités calculés avec confiance. Le premier reprend l'analyse faite à une qualité de lecture Q où les 12 bases du CDR3 ont en moyenne moins d'une chance sur mille d'avoir été mal lues (Q>30) Tab.2 et l'autre Tab.3, montre les ajustements des données sans se soucier de la qualité de lecture. Ce qu'il faut retenir de ces comparatifs est que dans la grande majorité des cas, changer le seuil pour les sélectivités bien estimées ne change pas le classe d'universalité de l'ajustement (à l'exception de NoFramework noire Q>0). Il semble même que dans les barres d'erreurs les différents estimateurs se recoupent. Si on compare l'effet de la qualité de lecture, dans la plupart des cas la classe d'universalité est aussi conservée. Ce n'est notablement pas le cas pour Frog3 PVP pour n>10 : pour les autres cas la P-value ne remet pas en cause le caractère exponentiel de la distribution mais on ne peut pas nier la différence en terme de comportement de l'estimateur. C'est aussi plutôt mitigé pour Mix21 contre PVP, où un plateau dans l'estimation de κ n'est pas toujours observable. La vrai différence vient plutôt de la capacité à trouver un plateau pour l'estimateur $\hat{\kappa}$, et porte sur le nombre de points permettant l'analyse en valeurs extrêmes. L'analyse est globalement plutôt robuste, même si parfois il est difficile de s'assurer de la relation entre τ, κ et le seuil. L'analyse est donc robuste en terme de comportement des différents estimateurs vis à vis de sous groupes de données différents. Pour autant tout les sous groupes de données ne permettent pas de statuer aussi clairement les uns que les autres sur leur nature "statistique extrême" ou encore de différencier les différentes classes d'universa-

lités rencontrées.

	Q>30		
	n>10	$\frac{\Delta s}{s} < 0.3$	$\Delta s < s_{noise}$
Human2 noire	$\kappa = 0.5 \pm 0.2$ P<10^{-4}	$\kappa = 0.75 \pm 0.35$ P<10^{-4}	$\kappa = 0.9 \pm 0.5$ P<10^{-4}
NoFramework noire	$\kappa = 0.03 \pm 0.08$ P=0.73	Pas assez de points	$\kappa = -0.1 \pm 0.12$ P=0.86
Frog3 PVP	$\kappa = 0.07 \pm 0.1$ P=0.97	$\kappa = -0.01 \pm 0.1$ P=0.94	$\kappa = 0.08 \pm 0.17$ P=0.63
Frog3 noire	Pas assez de points	Pas assez de points	Pas de plateau
NoShark PVP	$\kappa = 0.28 \pm 0.14$ P<10^{-4}	$\kappa = 0.27 \pm 0.19$ P=10^{-4} *	$\kappa = 0.16 \pm 0.11$ P=$2 * 10^{-4}$
Mix21 PVP	$\kappa = 0.42 \pm 0.22$ P<10^{-4}	$\kappa = 0.33 \pm 0.31$ P=$4.3 * 10^{-3}$	$\kappa = 0.46 \pm 0.29$ P=$2 * 10^{-4}$
Mix21 noire	$\kappa = -0.38 \pm 0.21$ P=$4.5 * 10^{-2}$	Pas assez de points	Pas assez de points
Mix 24 noire	Pas assez de points	Pas assez de points	Pas assez de points

TABLE 2: **Estimation de $\hat{\kappa}$ pour différents seuil de confiance sur la sélectivité et une qualité de lecture Q>30.** L'étoile pointe des estimateurs pour qui la relation entre τ, κ et le seuil n'est pas évidente.

	Q>0		
	n>10	$\frac{\Delta s}{s} < 0.3$	$\Delta s < s_{noise}$
Human2 noire	$\kappa = 0.5 \pm 0.2$ P<10^{-4}	$\kappa = 0.44 \pm 0.14$ P<10^{-4}	$\kappa = 0.45 \pm 0.2$ P<10^{-4}
NoFramework noire	Pas de plateau	$\kappa = 0.01 \pm 0.01$ P=0.93	$\kappa = -0.25 \pm 0.2$ P<10^{-4}
Frog3 PVP	$\kappa = 0.39 \pm 0.17$ P<10^{-4}	$\kappa = 0.15 \pm 0.12$ P<$3.2 * 10^{-2}$ *	$\kappa = 0.11 \pm 0.14$ P<0.2
Frog3 noire	$\kappa = -0.22 \pm 0.18$ P<0.22 *	Pas assez de points	Pas assez de points
NoShark PVP	$\kappa = 0.62 \pm 0.26$ P<10^{-4}	$\kappa = 0.6 \pm 0.2$ P<10^{-4}	$\kappa = 0.4 \pm 0.2$ P<10^{-4}
Mix21 PVP	Pas de plateau	Pas de plateau	Pas de plateau
Mix21 noire	$\kappa = -0.3 \pm 0.1$ P<10^{-4}	Pas assez de points	$\kappa = -0.28 \pm 0.13$ P=10^{-4} P<10^{-4}
Mix 24 noire	$\kappa = 0.14 \pm 0.1$	$\kappa = 0.18 \pm 0.19$ P=$7 * 10^{-3}$	$\kappa = 0.2 \pm 0.22$ P=0.1

TABLE 3: **Estimation de $\hat{\kappa}$ pour différents seuil de confiance sur la sélectivité et une qualité de lecture Q>0.** L'étoile pointe des estimateurs pour qui la relation entre τ, κ et le seuil n'est pas évidente.

En dehors de l'analyse en valeurs extrêmes, les fréquences et les corrélations ne sont pas sensibles aux erreurs de lectures.

Du point de vue de la sélectivité calculée entre les tour 1 et 2 ou entre les tours 2 et 3, on observe que les distributions et le comportement de l'estimateur $\hat{\kappa}$ sont très similaires Fig.51. Le coefficient de corrélation entre $s_{1,2}$ et $s_{2,3}$ quant à lui est très moyen mais montre quand même une corrélation significative par rapport au hasard : P= 10^{-86} calculé par le module Pearson corrélation de scipy. Pour pouvoir faire cette rapide analyse en terme de robustesse vis à vis de la sélectivité entre les tours, il faut cette fois-ci prendre toutes les séquences présents aux 3 tours et au moins présentes à 10 exemplaires à chaque tour : c'est pourquoi les distributions sont différentes de celles présentées dans l'analyse ci-avant. En faisant cela on rate de bonnes séquences en terme de sélectivités qui ne sont pas assez présentes au tour 1. Cela peut aussi expliquer le coefficient de corrélation entre sélectivités calculés entre 1 et 2 et 2 et 3, un peu bas : on rate surement des bonnes sélectivités qui ne sont pas pris en compte car trop peu présentes au tour 1 ce qui change radicalement les rapports de forces entre les deux Z_t (les constantes de normalisation entre tours). Cela peut aussi expliquer pourquoi dans certains cas la relation entre κ, τ et le seuil s^* n'est pas respectée ou que l'on ne trouve pas de plateau pour κ (on est donc plus dans le régime GPD). J'aimerais souligner deux points importants. Tout d'abord, Frog3 PVP et NoFramework noire qui sont les deux distributions exponentielles que l'on trouve et dont l'ajustement change avec les conditions de seuil sur la confiance sur la sélectivité et sur la qualité de lecture, ne sont pas non plus très robustes du point de vue de ce test de robustesse Fig.53 Fig.55. Pour Frog3 PVP, encore une fois les P-value calculées ont tendance à dire que toutes ses distributions sont de type exponentielles (ou du moins que l'on ne peut pas dire qu'elles ne le

sont pas), pourtant on ne peut pas ne pas remarquer la stricte différence en terme de comportement de l'estimateur. Deuxièmement, si on relaxe trop la condition 10 exemplaires au tour 1(10 exemplaires au tour1 est une contrainte très sévère), on voit apparaitre des relations linéaires évidentes à l'œil entre log de sélectivité et log du rang Fig.56 (évidentes à l'oeil mais pas forcément statistiquement significatives) .

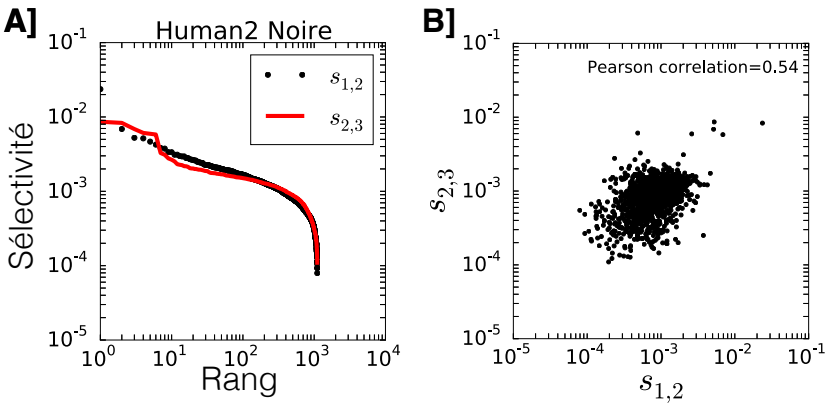


FIGURE 50: **Comparaison des distributions de sélectivités calculées entre différents tours.A]** Les distributions sont très similaires peu importe à partir de quels tours on les trace. **B]**Les sélectivités, elles, sont corrélées positivement, mais leur coefficient de corrélation est loin de 1, ce qui devrait être le cas si tout était absolument reproductible.

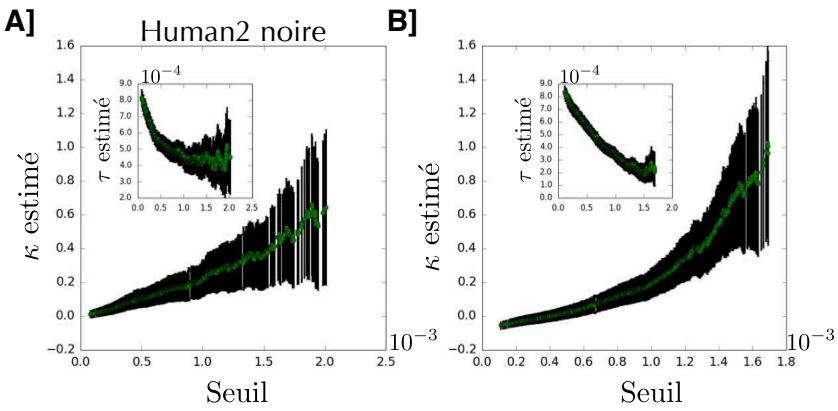


FIGURE 51: **Comparaison des ajustements de sélectivités calculées entre différents tours.A]** et **B]** Respectivement comportement de l'estimateur $\hat{\kappa}$ en fonction du seuil s^* pour les sélectivités calculées entre les tours1 et 2 et les tours 2 et 3. Ces comportements sont très similaires mais chose remarquable, aucune des deux courbes ne peut être correctement décrite par un GPD : ni plateau pour l'estimateur κ ni relation linéaire entre le seuil, τ et κ.

Vérification de la présence d'anticorps capables de se lier dans la sélection

Nous avons procédé à un test ELISA pour la présence dans le résultat de la sélection au tour 3, d'anticorps capables de s'attacher aux différentes cibles. Il s'agit de tests ELISA qualitatifs, présentés ici dans le seul but de montrer que la sélection a bien été dirigée contre les différentes cibles. : ce sont des données préliminaires. Des dots blot (technique permettant de quantifier l'expression et la sécrétion d'une protéine) ont été entrepris (données non disponibles) et ont montré que la luminescence de certains clones dans les expériences

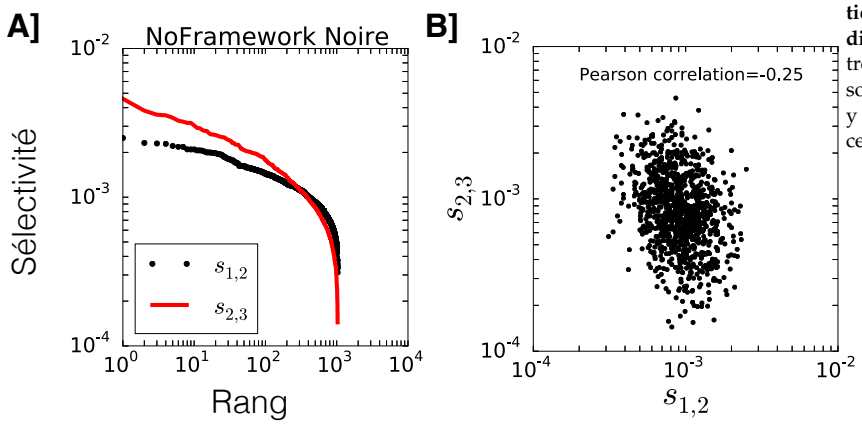


FIGURE 52: **Comparaison des distributions de sélectivités calculées entre différents tours.** Les distributions sont très peu similaires. Les sélectivités elles, sont anti-corrélées ce qui montre qu'il y a définitivement un problème dans cette sélection.

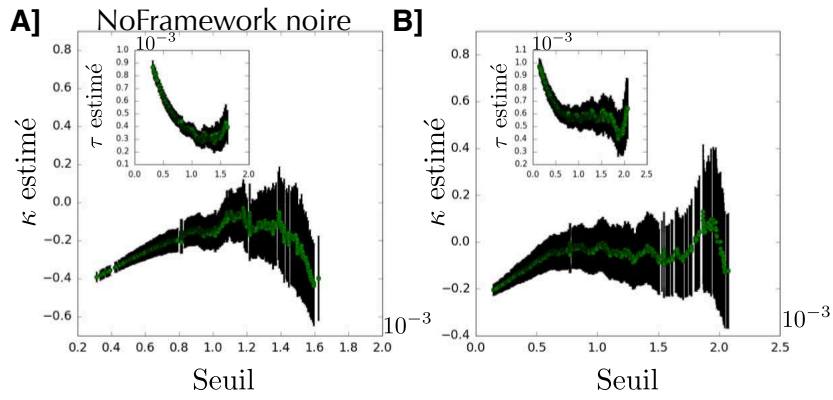


FIGURE 53: **Comparaison des ajustements de la distribution de sélectivités calculées entre différents tours. A]** et **B]** Respectivement comportement de l'estimateur $\hat{\kappa}$ en fonction du seuil s^* pour les sélectivités calculées entre les tours1 et 2 et les tours 2 et 3. Entre le tour 1 et 2 il est difficile de trouver un plateau dans le plot de κ vs le seuil. Entre le tour 2 et 3 la variation de l'estimateur de κ et de τ concorde pour dire que la distribution se comporte comme une exponentielle. En terme de comportement pur et dur de l'estimateur, sans parler de plateau, les courbes **A]** et **B]** se ressemblent au sens où elles ont tendance à identifier des exponentielles.

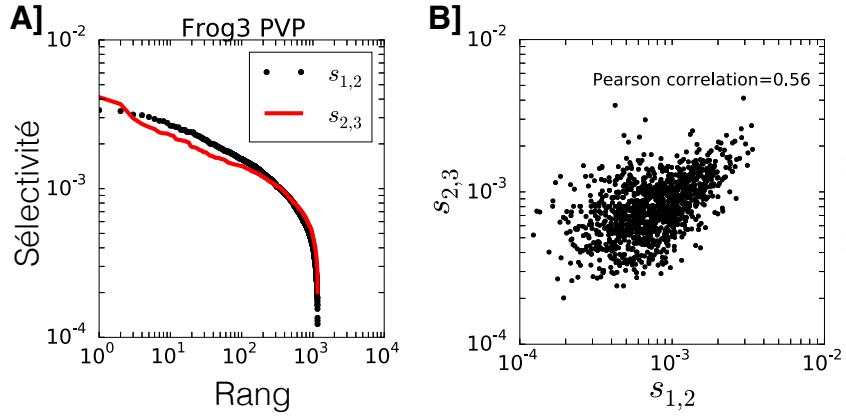


FIGURE 54: **Comparaison des distributions de sélectivité calculées entre différents tours. A]**Les distributions sont très peu similaires : à l'œil on voit que dans un cas le graphe est courbe alors que dans l'autre il dessine une ligne. **B]**Les sélectivités, elles, sont corrélées positivement, mais leur coefficient de corrélation est loin de 1, ce qui devrait être le cas si tout était absolument reproductible

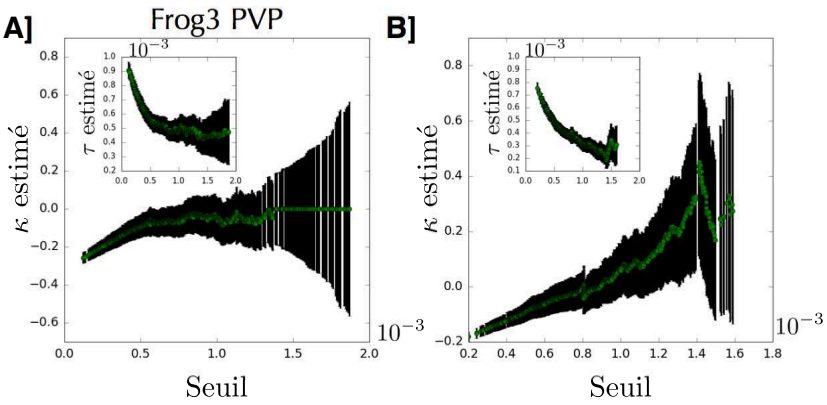


FIGURE 55: **Comparaison des ajustements de distribution de sélectivité calculées entre différents tours. A]** et **B]** Respectivement comportement de l'estimateur $\hat{\kappa}$ en fonction du seuil s^* pour les sélectivités calculées entre les tours 1 et 2 et les tours 2 et 3. Ces comportements sont très peu similaires. On a entre le tour 1 et 2 $\hat{\kappa} = -0,06 \pm 0.07$ avec $P = 0.28$ et entre le tour 2 et 3 l'impossibilité de déterminer un plateau pour l'estimation de κ. Même si on y tenait vraiment on se rend compte que la relation entre τ et le seuil n'est pas respecté.

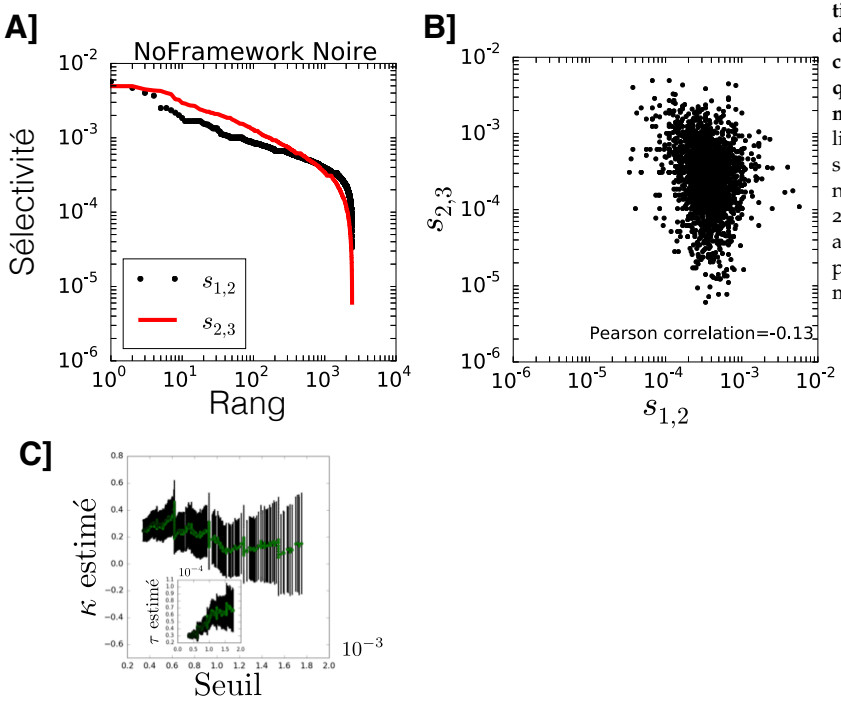


FIGURE 56: **Comparaison des distributions de sélectivités calculées entre différents tours dans le cas où aucune contrainte n'est mise sur les séquences que l'on considère comme bien estimée. A]** Les deux distributions sembles linéaires en log-log.**B]** Les sélectivités sont encore anti-corrélées.**C]** L'analyse n'a pas pu être faite entre le tour 1 et 2. Le tour2 et 3 donne un $\hat{\kappa} = 0,2 \pm 0.2$ avec $P = 0.08$ ce qui n'est pas suffisant pour réfuter l'hypothèse exponentielle malgré la belle linéarité vue en **A]**.

ELISA n'était pas seulement du à leur expression plus grande. En effet, un test ELISA est basé sur la mesure de la luminescence produite par des anticorps qui ont réussis à s'attacher à la cible qu'on leur présente et cela après des étapes de lavage. L'intensité de cette luminescence est directement liée au nombre initial d'anticorps présents dans le puits. Des différences significatives entre ces nombres initiaux peuvent apparaitre si les anticorps ne sont pas sécrétés au même taux. Celui-ci est en faite une convolution de l'efficacité qu'a l'anticorps de se lier à la cible mais aussi du nombre initiale d'anticorps présents. Pour être quantitatifs ces tests ELISA ont besoin d'être reproduits, analysés avec un logiciel de calcul d'intensités lumineuses et fait d'une manière plus systématique.
Dans tous les cas, sauf NoFramework contre la cible rouge, on retrouve entre 15 et 20% d'anticorps capables de se lier significativement à la cible. Pour Mix24 contre noire on a même 100% d'anticorps positifs. Mix24 contre la cible noire est une exception expérimentale : de manière reproductible (2 répliques) on a observé qu'à chaque tour de sélection la sélectivité globale gagnait un facteur 100.

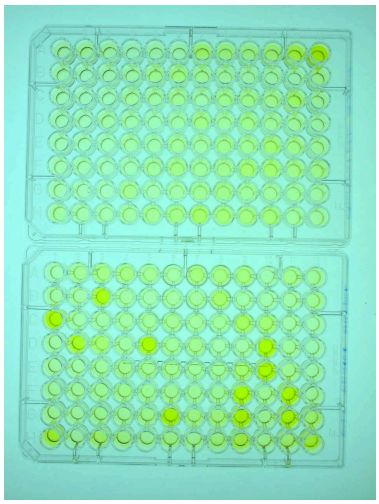


FIGURE 57: **Résultat d'une expérience ELISA cible PVP sur 96 anticorps tirés au hasard dans le résultat du tour 3 de sélection de Mix21 contre PVP. La plaque 96 puits du haut** n'a pas ses puits recouverts de PVP et représente la luminosité basale du bruit d'accroche non-spécifique de l'anticorps ou pour certains puits dont la luminosité est potentiellement haut dessus de la moyenne, l'accroche d'un anticorps spécifiquement au plastique. **La plaque 96 puits du bas** est recouverte de PVP. Certains puits se détachent clairement par leurs intensités, signe de la présence d'anticorps capable de lier spécifiquement au PVP. La comparaison des deux plaques permet de définir un seuil, pour établir qu'un signal est significatif ainsi que de s'assurer que la sélection a bien lieu contre le PVP et non le plastique.

Conclusions

Nous avons réussi à sélectionner des anticorps pour leurs capacités à se lier à une cible ADN (la cible noire) et un polymère (le PVP), à partir d'une banque d'anticorps très minimaliste. Nous avons pu confirmer l'apparition de corrélations entre les différents sites du CDR3, dues à la sélection. Les corrélations observées sont différentes des corrélations inhérentes au biais initiale des banques, et leur schéma est cible et banque dépendant. Des corrélations entre CDR3 et frameworks sont aussi observables, car différentes banques présentent différentes "séquences consensus" pour la même cible : Human2 contre noire ("QQQQ"), Frog3 contre noire ("AELD"), human germline contre noire ("RRKH"), chicken contre noire ("GRFP ") (analyse au cas par cas en annexe). De plus l'expérience de phage

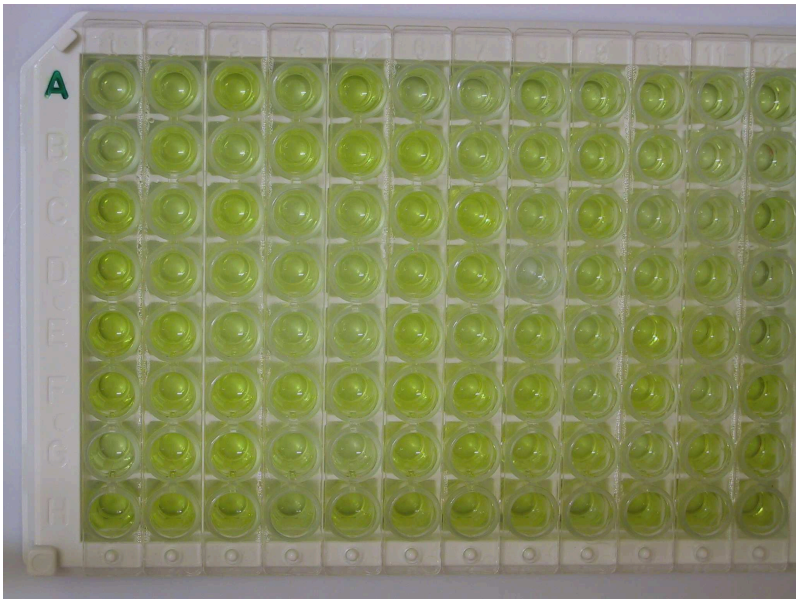


FIGURE 58: **Résultat d'une expérience ELISA cible noire sur 96 anticorps tirés au hasard dans le résultat du tour 2 et 3 de sélection de Mix24 contre noire.** Les rangées A à D sont des anticorps issus du tour 2 et ceux des rangées E à H sont issus du 3ème tour. Ils semblent tous être capables de bien se lier à la cible. Ce résultat peut paraitre étonnant mais Mix24 contre noire est une expérience exceptionnelle : à chaque tour la sélectivité global gagnait un facteur 100.

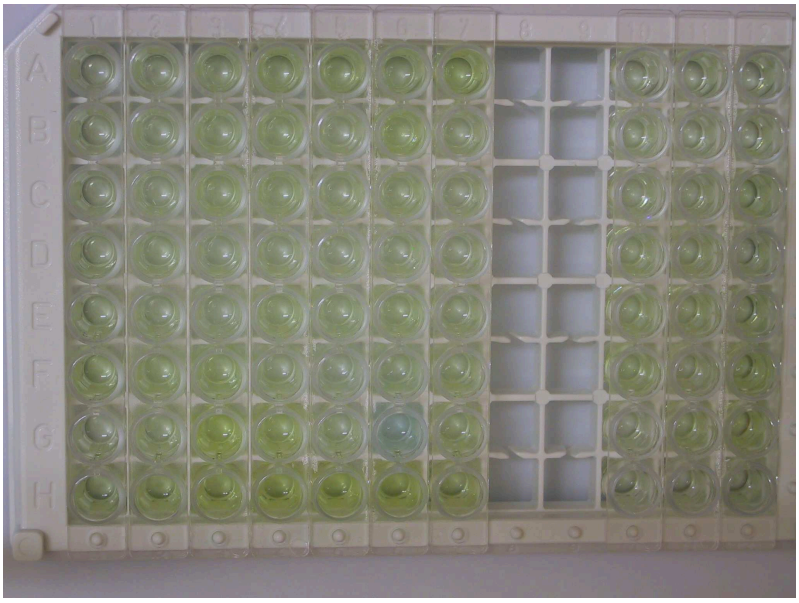


FIGURE 59: **Résultat d'une expérience ELISA sans cible noire sur 96 anticorps tirés au hasard dans le résultat des tours 2 et 3 de sélection de Mix24 contre la cible noire.** Les rangées A à D sont des anticorps issus du tour 2 and ceux des rangées E à H sont issus du 3 ème tour. Les puits H11 et H12 sont sans anticorps. Cette fois on voit qu'il y a très peu d'affinité comparé à Fig.58, pour la plaque simplement recouverte de streptavidine et non de cibles. Pour autant on peut deviner des anticorps plus ou moins significativement capable de se lier à la streptavidine, apparaitre au tour 3.

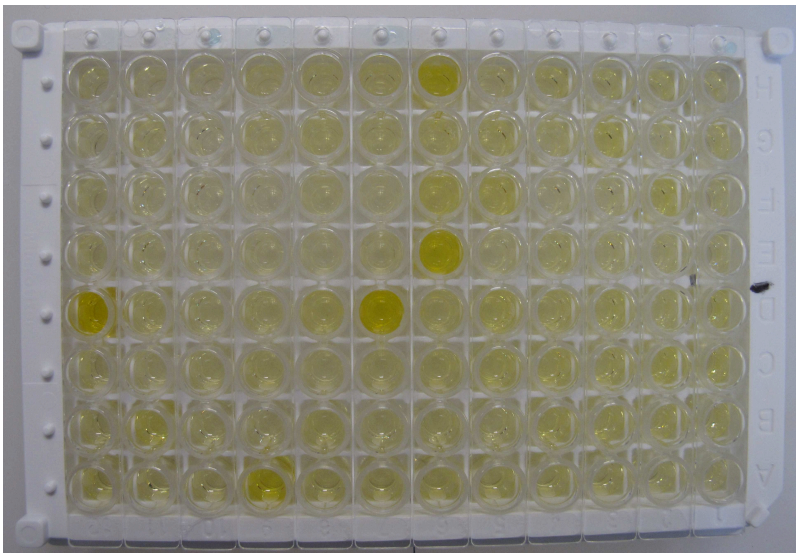


FIGURE 60: **Résultat d'une expérience ELISA rouge sur 45 anticorps tirés au hasard dans le résultat du tour 3 de sélection de Mix24 contre la cible rouge.** Les rangées A à D sont munis de la cible rouge et les rangées E à H sont sans cible rouge. A1 et E1 sont sans anticorps. Encore une fois il y a une forte sélection pour les anticorps capables de s'attacher spécifiquement à la cible rouge. Certains anticorps anti streptavidine existent pour autant : H6 et E6

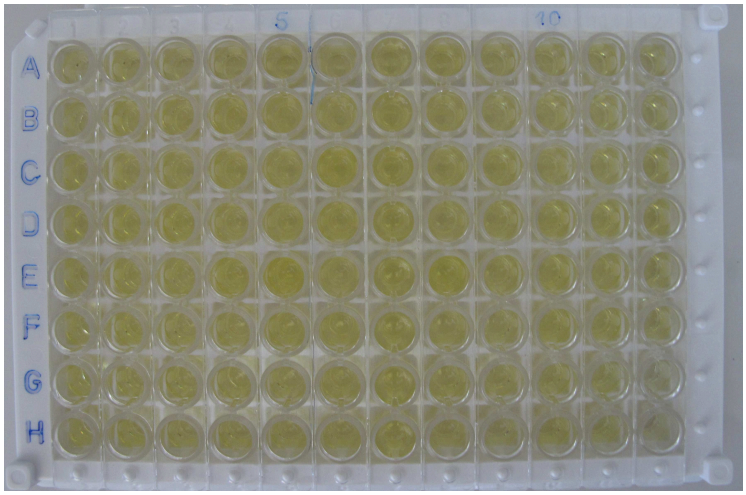


FIGURE 61: **Résultat d'une expérience ELISA rouge sur 45 anticorps tirés au hasard dans le résultat du tour 3 de sélection de NoFramework contre la cible rouge.** Les rangées A à D et E et H sont des répliques : A=E, B=F etc. H12 et D12 sont sans anticorps. Il n'y pas d'intensités significativement plus importantes que la luminescence basale.

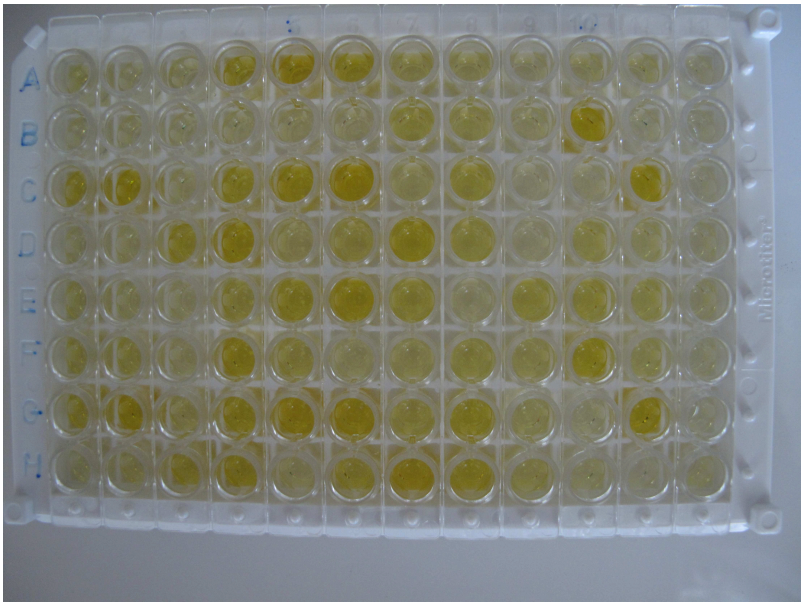


FIGURE 62: **Résultat d'une expérience ELISA cible noire sur 45 anticorps tirés au hasard dans le résultat du tour 3 de sélection de Frog3 contre la cible noire.**Les rangées A à D et E et H sont des répliques : A=E, B=F etc. H12 et D12 sont sans anticorps. Des anticorps anti cible noire sont bien présents. On voit que la mesure ELISA dans ce cas est plutôt variable.

display est reproductible.
L'analyse en terme de valeurs extrêmes a montré sa cohérence et sa faisabilité. Je pense en effet qu'il s'agit là de la première analyse en valeurs extrêmes portant sur l'évolution, et utilisant les valeurs extrêmes avec un jeu de données aussi important. L'hypothèse portant sur le fait que l'on soit sur la queue de la distribution n'est pas totalement absurde pour deux raisons. La première déjà énoncée, est que le CDR3 est connu pour être l'ensemble de sites de l'anticorps ayant le plus d'impact sur la liaison. La seconde, plus pratique mais tout aussi qualitative, est que l'on retrouve bel et bien des plateaux dans l'estimation de κ en fonction du seuil (seul bémol, cela est parfois dépendant jeux de données). De plus les ajustements sont dans l'ensemble plutôt corrects et représentent donc, je pense, au moins une bonne approximation dans la description d'un processus de sélection. Un résultat important étant que l'on peut a priori retrouver les 3 classes d'universalité. N'oublions pourtant pas l'incapacité d'ajuster Frog 3 contre noire.
La robustesse de l'analyse dans le régime de l'expérience a été établie. Je pense avoir montré que dans le régime de séquençage dans lequel j'ai travaillé, les incertitudes sur l'estimation de sélectivités jouaient un rôle minime dans l'analyse en valeurs extrêmes. Dans le cas des erreurs de lectures, l'analyse est plutôt robuste. En effet si on décide d'occulter NoFramework noire qui semble avoir été sélectionné bizarrement (voire anti-corrélations des sélectivités entre tour1 et 2 et 2et 3), le seul autre cas qui peut poser problème est Frog3 PVP. Dans ce cas, deux des trois P-value indiquent bien que l'on ne peut pas faire la différence entre les différentes classes d'universalités aux vues des données : ainsi Q>0 et Q>30 ne se contredisent pas. Pour la troisième P-value, je n'ai pas d'explications.

Trois points peuvent être améliorés dans cette expérience :

- La qualité de lecture, notamment en rapprochant le début de la lecture au niveau du CDR3 en désignant de nouveaux primers .
- Le nombre de fragments lus : multiplexer moins d'expériences, se focaliser sur le séquençage de 2 round consécutifs choisis, et bien choisir sa technologie de séquençage (ici Hiseq plutôt que Miseq).
- Avoir accès à une gamme dynamique des sélectivités plus grande : soit en partant de banques initiales plus homogènes (en pratique durement faisable), soit en sélectionnant mieux les tours à séquence (moins efficace et demande de l'expérience.)

Primer : Petite séquence ADN complémentaire spécifique d'un brin d'ADN simple que l'on veut amplifier et qui servira de support à l'ADN polymérase pour la reconstruction du brin double.

Le troisième point renforce le fait que toutes les sélections ne se passent pas de la même manière : certaines sont très efficaces (les anticorps sont bons en terme de liaison à la cible) et séquencer le tour 1 et la banque initiale pourrait suffire (on garde ainsi bien plus de diversités). D'autres sont très lentes et il aurait fallu faire des tours en plus (pour réellement voir des coefficients de sélectivités significatifs apparaitre), d'autres encore de par la nature des distributions de sélectivités impliquent des écarts énormes entre mutants et réduisent très vite la diversité observable par séquençage.
Les deux premiers points sont en ce moment même traités car nous reséquençons en prenant en compte ces améliorations.
Enfin, dans une dernière partie, j'aimerais ouvrir mes résultats et ce set up expérimental, à d'autres expériences soit déjà entreprises soit à proposer.

Perspectives

Prochaines expériences

Comme énoncé précédemment, certaines expériences vont être reséquencées. Le fait est que plus de trente sélections (fois trois tours de sélections) ont été entreprises au cours de la thèse et celles qui ont été séquencées jusqu'à aujourd'hui en représente une petite partie. Nous sommes en ce moment en train de préparer des nouveaux échantillons à séquencer pour étayer l'analyse.

Réponse à la sélection de banques aux frameworks différenciés par quelques mutations.

Ceux-ci concernent principalement les frameworks Human germline, Human limited et Human Broadly Neutralizing (BNAB). En effet, des sélections de ces trois frameworks contre les trois cibles ADN ont été faites. Il s'agira de séquencer les résultats de sélections de ces 3 frameworks contre la cible noire et bleue. Le but est de comparer la sélection de 3 banques ayant des frameworks généalogiquement liés et un nombre de mutations moyen et intermédiaire par rapport au germinal. Cette étude est poussée par le fait que des différences flagrantes de sélection ont été observées pour ces 3 frameworks contre les deux cibles différentes lors de nos expériences. En effet la sélectivité globale, c'est à dire le ratio globale du nombre de phages avant et après sélection pour les deux cibles, suit sans équivoque la relation suivante : $S_{germinal} >> S_{Limited} >> S_{BNAB}$. Human2 contre bleue et Frog3 contre bleue seront aussi séquencés pour compléter les mesures déjà entreprises contre la cible noire.

Réponse à la sélection d'un framework germinal muté.

Une deuxième expérience déjà en cours et rendue possible grâce aux résultats de cette thèse, est celle portant sur l'ajout contrôlé de mutations dans un framework et de la comparaison de la capacité de sélection et d'evolvabilité qui en résulte. En effet une banque de framework Human germline muté a déjà été créée. La sélection pour un framework muté contre la cible noire nous apportera une idée de la statistique des effets de sélectivité induite par des mutations dans le framework et nous permettra de comparer la capacité de sélection d'une banque de CDR3 formé à partir de ce framework muté pour noire, lors de sélection contre noire et bleue.

Evolvabilité : Trait qui caractérise la capacité d'un mutant à évoluer.

Etudes préliminaires qualitatives à pousser

J'ai procédé à deux autres types d'études qui étant donné la qualité des données n'ont pas rendu possible une analyse quantitative :

- Etude du shape space : en s'intéressant notamment à reconnaitre des anticorps communs à plusieurs cibles.
- Etude du fitness landscape locale accessible par l'expérience.

Pour autant je pense que ce sont des questions intéressantes et accessibles avec le set up expérimental utilisé dans cette thèse.

Shape space

Shape space : Utilisé principalement dans l'étude théorique des anticorps, il s'agit d'un espace de cibles où la notion de distance, bien que bancale, est définie par l'affinité relative d'un anticorps pour deux cibles.

En sélectionnant une banque contre plusieurs cibles on peut se demander quelle est la probabilité de trouver un anticorps capable de lier à plus d'une cible ? Où se placent ces anticorps par rapport aux autres anticorps sélectionnés ? Ce genre de données peuvent permettre de reconstruire un shape space, notamment en appliquant une méthode de multidimensional scalling [61] [62] [63] . On a ainsi accès à la dimension de cet espace ainsi qu'à une façon d'ordonner des anticorps pour par exemple former un chemin continu entre deux cibles. Avoir une idée de la dimension du shape space permet de redéfinir la diversité nécessaire d'un répertoire immunitaire naturel pour que celui-ci soit capable de répondre à n'importe quel problème de liaison à un antigène.

Nous n'avons pas assez de données pour répondre à ces questions. Pour autant nos expériences permettent déjà de s'apercevoir que trouver des anticorps identiques dans deux sélections différentes n'est pas si difficile que cela.

Le framework chicken se retrouve en quantité suffisante pour une analyse contre trois cibles : noire, bleue et rouge. Pour ces trois cibles la même séquence consensus est sélectionnée : 'GR(MF)P' (voir annexe), ce qui est assez troublant et indique peut être une sélection pour la tige plus que la boucle de la cible d'ADN. On retrouve quelques séquences partagées entre ces trois cibles Fig.63

Human germline offre aussi quelques séquences communes à la cible noire et bleue Fig.64 : cette fois les séquences consensus sont différentes. Enfin on peut faire de même pour Frog3 contre noire et PVP Fig.65 : on peut supposer que ces deux cibles sont très éloignées par rapport aux cibles ADN entre elles.

61. Eleni Katifori David Jordan, Seppe Kuehn and Stanislas Leibler. Behavioral diversity in microbes and low dimensional phenotypic spaces. *Proc. Natl. Acad. Sci. USA*, Early edition:1–6, 2013

62. Derek J. Smith et al. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305:371–375, 2004

63. Alan Lapedes and Robert Farber. The geometry of shape space : application to influenza. *Santa Fe Institute*

Multidimensional scalling : Méthode, qui, étant donné la seul donnée des distances entre points, permet de retrouver les coordonnées de ceux-ci.

64. William Bialek Thierry Mora, Aleksandra M. Walczak and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA*, 107:5405–5410, 2010

65. Roonen Segev Elad Schneidman, Michael J. Berry and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007–1012, 2006

66. Mehran Kardar Jhon P. Barton and Arup K. Chakraborty. Scaling laws describe memories of host-pathogen riposte in the hiv population. *Proc. Natl. Acad. Sci. USA*, 112:1965–1970, 2014

Fitness landscape

Une autre analyse intéressante possiblement réalisable avec ce genre de données est la reconstitution d'un fitness landscape : combien de maximum locaux, à quel point sont-ils pentus, combien de chemins permettent d'y accéder ? Quelle est la distribution des sélectivités à un, deux ou trois points de mutations des différents maximums. Cette analyse serait réalisable en traitant les données de sélectivité selon un modèle de maximum entropy [64] [65] [66]. L'analyse stricto

Maximum entropy model : Il s'agit de représenter une connaissance imparfaite d'un phénomène par une distribution de probabilité en choisissant celle ci avec le moins d'aprioris possible. En ce sens cette distribution doit maximiser l'entropie.

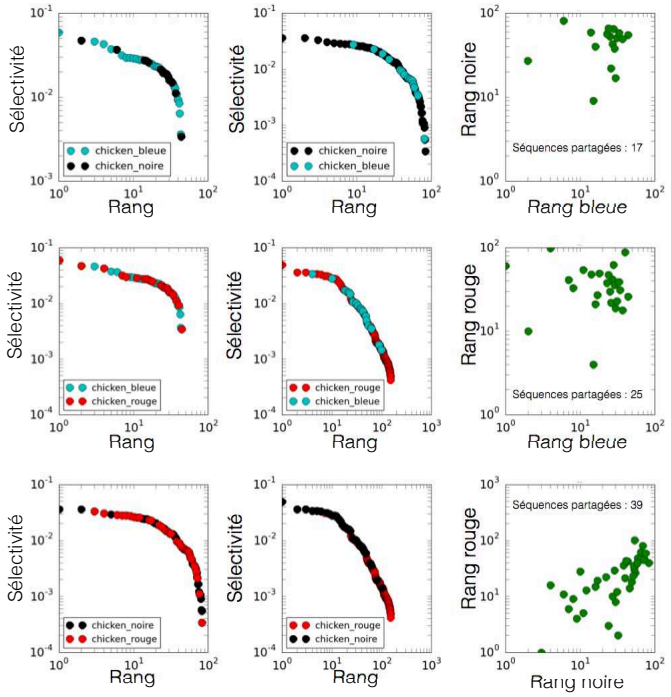


FIGURE 63: **Rang des séquences partagées pour chaque paire de cibles pour la banque chicken.** Les nombres de séquences sont équivalents pour les 3 paires, lorsque pondérés par la taille des différentes diversités. Les séquences sont uniformément répartis en terme de rang dans les différentes distributions. Il s'agit très certainement d'une sélection pour la partie commune de la cible, sans pour autant que l'on puisse en être sûr.

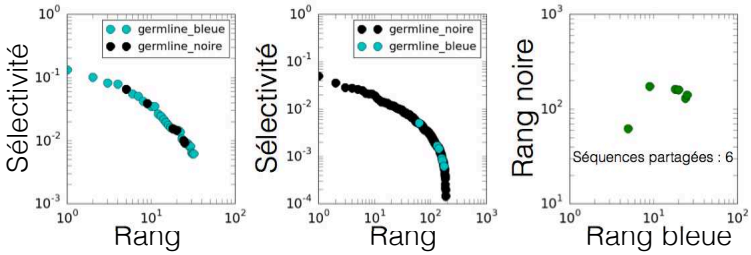


FIGURE 64: **Rang des séquences partagées pour la paire de cibles noire/bleue pour la banque Human Germline.** 6 séquences sont partagées. Elles sont assez bien répartie en terme de rang chez germline contre la cible bleue, mais sont clairement à bas rang chez germline contre la cible noire.

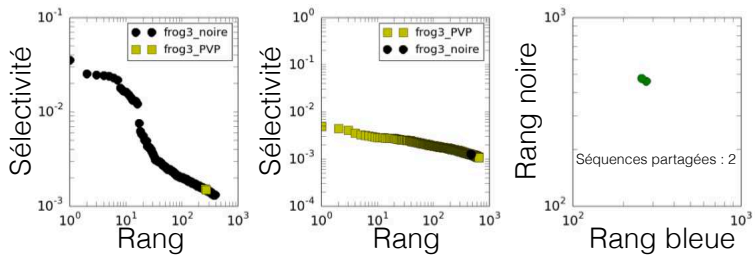


FIGURE 65: **Rang des séquences partagées pour la paire de cibles noire/PVP pour la banque Frog3.** 2 séquences seulement sont partagées entre ces deux cibles potentiellement très éloignées de par leur nature. Les deux séquences se situent parmi les mauvais lieurs pour les deux cibles.

sensu du fitness landscape par un modèle d'entropie maximum n'a pas pu être réalisée : cette analyse demande, dans notre cas, d'ajuster plus d'un millier de paramètres si on veut un modèle de corrélations à deux sites. Aucune des expériences ne présente les deux conditions requises pour lancer cette étude : une grande confiance dans l'estimation des fréquences et des différences de fréquences ainsi que plusieurs milliers de points. Mix21 contre PVP est une expérience où la confiance sur les sélectivités est plutôt bonne, par contre la distribution de sélectivités retenue n'a que 400 séquences.
On peut tout de même regarder comment les anticorps sont répartis dans la distribution de sélectivités en fonction de leur distance de Hamming . Prenons une séquence dont la sélectivité est bien évaluée, la séquence de rang 4, et regardons combien de voisins à 1, 2, 3 ou 4 points de mutations elle a, et comment les sélectivités correspondantes sont distribuées entre elles Fig.66 et à l'intérieur de la distribution totale Fig.67 .
Enfin, je pense qu'il serait intéressant de regarder le lien entre fitness landscape (par le biais du modèle d'entropie maximum) et les différentes classes d'universalités du théorème des valeurs extrêmes : quel type de corrélations induit quel type de classes d'universalités, pour que potentiellement κ soit porteur d'une information supplémentaire de type topographique. il est aussi intéressant de noter que l'on retrouve cette loi de puissance pour des diversités de banques réduites Fig.66 : si on autorise q'un point de mutation, strictement 2 ou 3 ou 4. Ce genre d'analyse pourrait donc servir à se donner une idée de la diversité à mettre dans une banque pour pouvoir atteindre un certain degré performance pour les meilleurs mutants : si on multiplie la diversité par 10 on multiplie la performance du meilleur mutant par κ.

Distance de Hamming : La distance de Hamming est définie par le nombre de différences entre deux séquences. Par exemple, AAG et GAA ont une distance de Hamming de 2.

Dynamique de sélection.
Il est intéressant de remarquer que si la banque initiale était parfaitement uniforme, l'abondance des séquences après un tour de sélection devrait être décrite par la distribution de sélectivités. Un second tour reviendrait, en terme de fréquences, à passer chaque sélectivité au carré, un troisième tour à les mettre au cube et ainsi de suite. Dans le cas d'une distribution de sélectivités de type Fréchet, $-\kappa$ représente la pente du log log plot sélectivité vs rang. Un second tour de sélection revient à multiplier cette pente par 2, mais cela revient aussi à augmenter l'écart entre deux sélectivités de rangs consécutifs : passant de $1+\kappa$ Fig.68 à Fig.70 $1+2\kappa$. Et ainsi de suite pour d'autres tours de sélection. Mais cela n'est pas vrai pour les autres classes d'universalité Fig.70 : l'écart entre deux sélectivités de rangs consécutifs ne varie pas linéairement avec n, où n est le nombre de tours de sélection. La relation est plus compliquée que cela, pour autant il semble que l'on puisse définir un $\kappa_{effectif}$ Fig.69 qui décrive ces différences de sélectivités. Ce $\kappa_{effectif}$ peut d'ailleurs être d'une tout autre classe d'universalité si on laisse la sélection opérer suffisamment longtemps. Pourtant, même avec un $\kappa_{effectif}$ positif définit

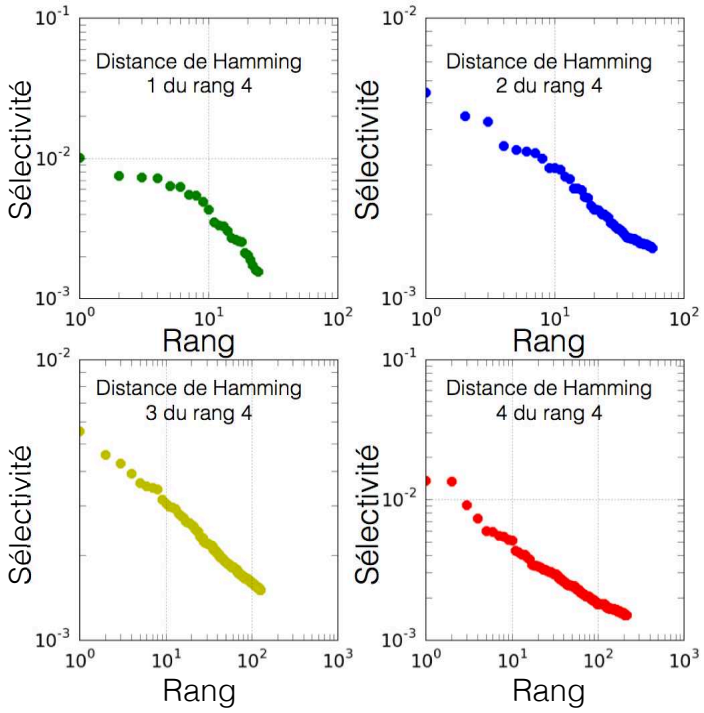


FIGURE 66: **Distribution de sélectivités de séquences à une distance 1, 2, 3 et 4 de la séquence de rang 4 dans la distribution de sélectivités totale.** Il est intéressant de remarquer que même en réduisant la diversité, c'est à dire en classifiant les séquences dont on regarde la sélectivité en fonction de leur distance par rapport à une séquence de référence, une relation linéaire est encore visible.

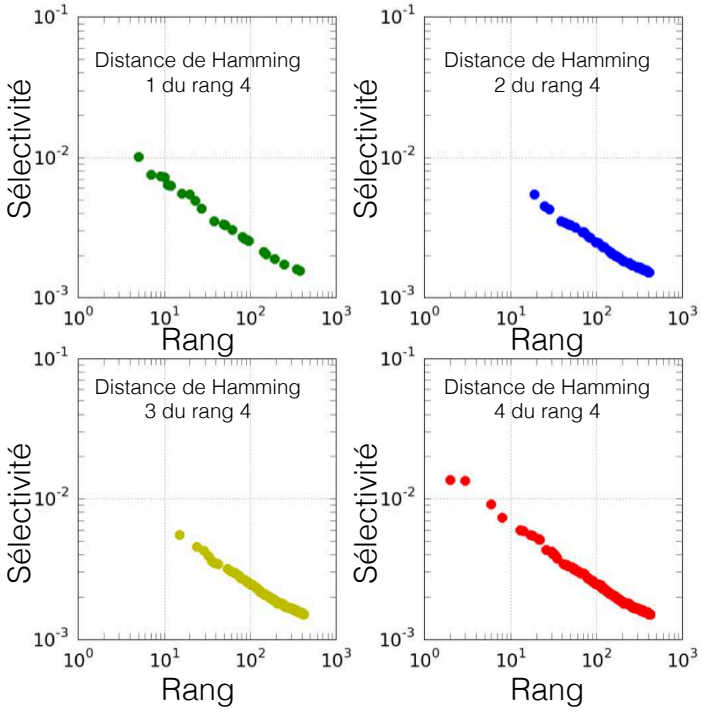


FIGURE 67: **Rang des séquences à une distance 1, 2, 3 et 4 de la séquence de rang 4 dans la distribution de sélectivités totale.** Les séquences à distance 1 se retrouvent d'une manière plutôt homogène dans toutes les décades de rang. Ceux à distances deux et trois ne sont pas dans le top 10. Au rang 2 se trouve une séquence à une distance de Hamming 4 de la séquences présente au maximum. Cela fait dire qu'il y a surement plusieurs maximum, et une étude de la distribution des sélectivités étant donné un maximum est surement très informative.

grâce aux différences de s , on ne retrouve pas une relation linéaire dans le log-log plot fréquences vs rang Fig.71 : différence de sélectivités et ratio de sélectivités se comportent différemment.
La sélection est un processus qui opère sur les différences entre sélectivités ou les ratios de fréquences. J'aimerais donc ajouter deux petites remarques pour conclure :

- Représenter les distributions de sélectivités en log-log n'est peut être pas une manière optimale de représenter ces distributions notamment si celle-ci ne sont pas de type Fréchet.
- Nos résultats sont compatibles avec les lois de puissances trouvées dans les répertoires immunitaires[67]. Cela implique par contre que les distributions de sélectivités soient au départ de type Fréchet. La pente peut alors être interprétée comme étant le niveau de sélection auquel la banque a déjà fait face.

Ainsi cette expérience qui s'est focalisée sur la caractérisation de la sélection dans des banques d'anticorps, peut servir de point de départ du point de vue expérimentale et de l'analyse, pour l'exploration de nombreux autres sujets au coeur des thématiques contemporaines de la recherche dans de nombreux domaines : de la compréhension des événements extrêmes à la conception de banques de protéines, en passant par l'immunologie.

67. William Bialek Thierry Mora, Aleksandra M. Walczak and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA*, 107:5405–5410, 2010

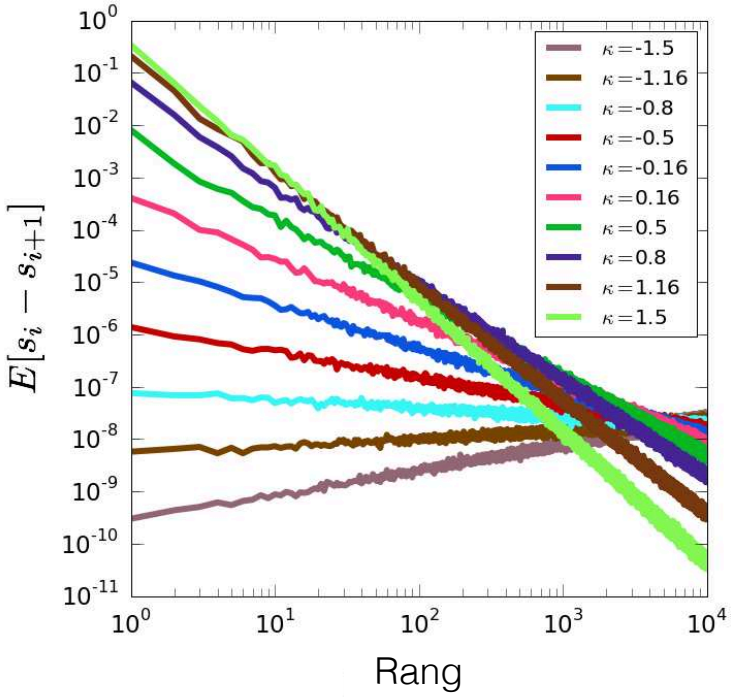


FIGURE 68: **Simulations. Relation entre le paramètre d'universalité κ et les écarts de sélectivités.** Des simulations montrent que κ est directement relié à l'espacement moyen en terme de sélectivité entre deux mutants adjacents en termes de rang : $\log(E[\Delta_i]) = -(\kappa + 1) \times \log(i) + \log(E[\Delta_1])$ où $\Delta_i = s_i - s_{i+1}$, E[] la moyenne sur plusieurs tirages, et i le rang.

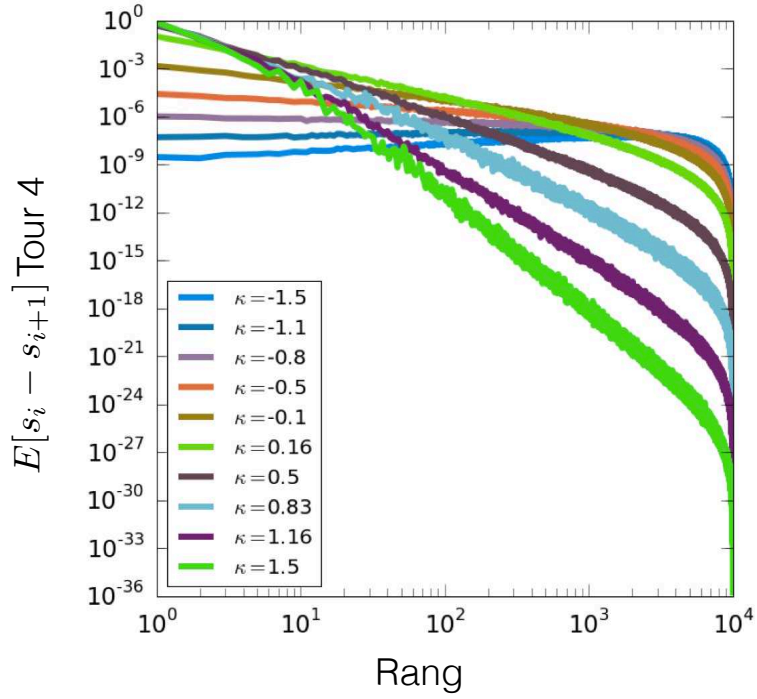


FIGURE 69: **Simulations. Relation entre le paramètre d'universalité κ et les écarts de sélectivités après 4 tours de sélection.** Comme pour la Fig.16, au moins pour les 500 premiers points, on peut voire une relation linéaire dans le log log plot de $E[\Delta_i]$ vs rang dont la pente diffère avec le κ original. On peut donc en déduire un $\kappa_{effectif}$ représenté en Fig.70.

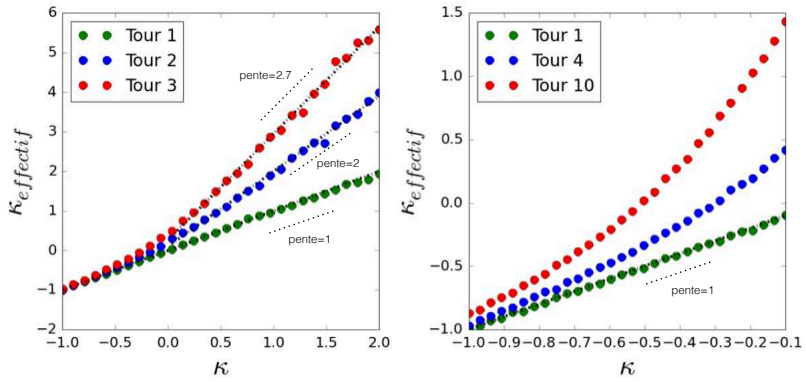


FIGURE 70: **Simulations. Relation entre κ et $\kappa_{effectif}$.** Les différentes classes d'universalité, ne se comportent pas de la même manière avec les tours de sélection. Pour les distributions où $\kappa > 0$ il semble y avoir une relation linéaire claire et tour dépendante, entre κ et $\kappa_{effectif}$. Pour les distributions où $\kappa < 0$, la relation n'est pas linéaire et permet même d'atteindre des $\kappa_{effectif} > 0$

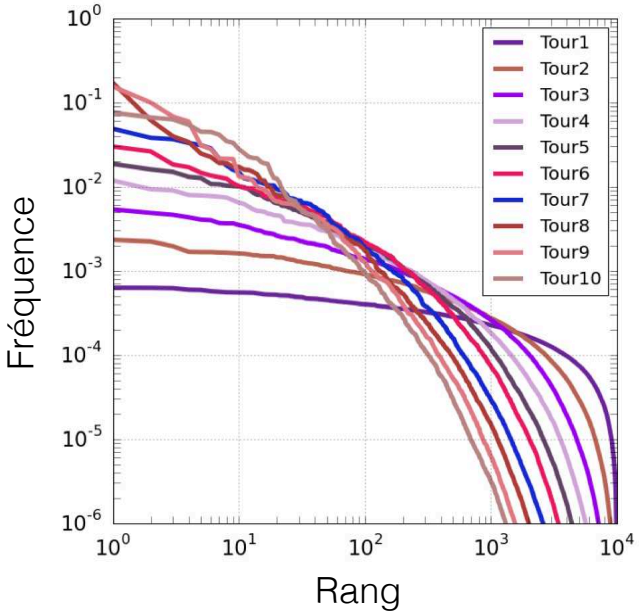


FIGURE 71: **Simulations. Evolution de la distribution des fréquences avec les tours de sélections pour une distribution initiale de sélectivités de $\kappa = -0.1$.** Avec un tel κ initiale, au tour 10 on pourrait s'attendre à voir un droite de pente 1.5 dans ce log log plot fréquence vs rang, si on se base sur le $\kappa_{effectif}$ caluclé en Fig.70. Ce n'est pas le cas, la courbe garde son caractère courbe initiale.

Bibliographie

[1] Charles Darwin. *On the Origin of species*. Oxford University Press, Oxford, 2008.

[2] Gregor Mendel. Experiments on plant hybridization. *Proceedings of the Natural History Society of Brunn*, 1866.

[3] W.S. Sutton. On the morphology of the chromosome group in brachystola magna. *Biol. Bull.*, 4 :24–39, 1902.

[4] W.S.l Sutton. The chromosomes in heredity. *Biol. Bull.*, 4 :231–251, 1903.

[5] T.H. Boveri. Ergebnisse uber die konstitution der chromatischen substanz des zelkerns. 1904.

[6] Oswald T.Avery and al. Studies on the chemical nature of the substance inducing transformation of pneumoccocal types. *J. Exp. Med.*, 79 :137–158, 1944.

[7] M. Meselson and al. The replication of dna in escherichia coli. *PNAS*, 44 :671–682, 1958.

[8] William B. Provine. *The origins of Theoritical Population Genectics*. The university of Chicago Press, 1992.

[9] Yule. Mendel s law and their probable relations to intraracial heredity. page 225.

[10] Howard C. Warren. Numerical effect of natural selection acting upon mendelian characters. *Genetics2*, pages 305–312, 1917.

[11] R.A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Royal Society of Edinburg*, 1918.

[12] R.A.Fisher. On the dominance ratio. *Proceeding of the Royal Society of Edinburgh*, 42 :321–341, 1922.

[13] H. Allen Orr. The genetic theory of adaptation : a brief history. *Nature review Genetics*, 6 :119–127, 2005.

[14] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge Univ. Press, 1983.

[15] Maynard Smith. *The Scientist Speculates : an Anthology of Partly-Baked Ideas*. Basic Books,New York, 1962.

[16] S. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, 128 :11–45, 1967.

[17] M. Kimura. Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl Acad. Sci. USA*, 76 :624–626, 1979.

[18] J. H. Gillespie. A simple stochastic gene substitution model. *Theor. Popul. Biol.*, 23 :202–215, 1983.

[19] J. H. Gillespie. Molecular evolution over the mutational landscape. *Evolution*, 38 :1116–1129, 1984.

[20] F. H. C. Crick J. D. Watson. Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid. *Nature*, 171 :737–738, 1953.

[21] Chargaff E Elson D. On the deoxyribonucleic acid content of sea urchin gametes. *Experientia*, 8 :143–145, 1952.

[22] M. Delbrück S. E. Luria. Mutation of bacteria from virus sensitivity to virus resistance. *Genetics*, 230 :491–510, 1943.

[23] Randall K. Saiki et al. Enzymatic amplification of beta globulin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *JSTOR*, 230 :1350–1354, 1985.

[24] Marshall W. Nirenberg et al. Ribonucleotide composition of the genetic code. *Biochemical and Biophysical Research Communications*, 6 :410–414, 1962.

[25] A. R. Coulson F. Sanger, S. Nicklen. Dna sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, 74 :5463–5467, 1977.

[26] Etienne Danchin and al. Beyond dna : integrating inclusive inheritance into an extended theory of evolution. *Nature Review Genetics*, 12 :475–486, 2011.

[27] Oded Rechavi and al. Transgenerational inheritance of an acquired small rna base based antiviral response in c. elegans. *Cell*, 147 :1248–1256, 2011.

[28] Simon D. Wagner and Michael S. Neuberger. Somatic hypermutation of immunoglobulin genes. *Annu. Rev. Immunol.*, 14 :441–457, 1996.

[29] Georges A. Gutman. Chapter 8 : genetic basis of antibody diversity. *Medical immunology 544*, Fall 2011.

[30] Anderw C. R. Martin Robert M. MacCallum and JAnet M. Thorton. Antibody-antigen interactions : contact analysis and binding site topography. *J. Mol. Biol.*, 262 :732–745, 1996.

[31] Marc M. Davis Jhon L. Xu. Diversity in the cdr3 region of vh is sufficient for most antibody specificities. *Immunity*, 13 :37–45, 2000.

[32] Florian Klein et al. Somatic mutations of the immunoglobulin framework are generally required for broad and potent hiv-1 neutralization. *Cell*, 153 :126–138, 2013.

[33] Feng Wan et al. Somatic hypermutation maintains antibody thermodynamic stability during affinity maturation. *PNAS*, 110 :4261–4266, 2013.

[34] Hennie R Hoogenboom. Selecting and screening recombinant antibody libraries. *Nature biotechnology*, 23 :1105–1116, 2005.

[35] G.P. Smith. Filamentous fusion phage : novel expression vectors that display cloned antigens on the vision surface. *Science*, 298 :1315–1317, 1985.

[36] J. et A. Pluckthun Hanes. In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. USA*, 94 :4937–4942, 1998.

[37] M.J. Feldhaus. Flow cytometric isolation of human antibodies from a non immune saccharomyces cerevisiae surface display library. *Proc. Natl. Acad. Sci. USA*, 94 :4937–4942, 1998.

[38] Douglas m Fowler et al. High-resolution mapping of protein sequence-function relationships. *Nature*, 7 :741–746, 2010.

[39] J. et al. McCafferty. Phage antibodies : filamentous phage displayng antibody variable domains. *Nature*, 348 :552–554, 1990.

[40] Modi S. et al. Two dna nanomachines map ph changes along intersecting endocytic pathways inside the same cell. *Nature Nanotechnology*, 8 :459–467, 2013.

[41] Soshee A. et al. General in vitro method to analyze the interactions of synthetic polymers with human antibody. *Biomacromolecules*, 15 :113–121, 2014.

[42] Frederic A. Fellous et al. Molecular recognition by binary code. *J. Mol. Biol.*, 348 :1153–1162, 2005.

[43] Frederic A. Fellous et al. High-throughput gene ration of synthetic antibodies from bighly functional minimalist phage displayed libraries. *J. Mol. Biol.*, 373 :924–940, 2007.

[44] Alan S. Perelson et Gerard Weisbuch. Immunology for physicists. *Review of modern physics*, 69 :1219–1262, 1997.

[45] Anthony D. Keefe et Jack W. Szostak. Functional proteins from a random-sequence library. *Nature*, 410 :715–717, 2001.

[46] Angus Buckling R. Craig Maclean. The distribution of fitness effects of beneficial mutations in pseudomonas aeruginosa. *Plos genetics*, 5, 2009.

[47] Darin R. Rokyta et al. Beneficial fitness effects are not exponential for two viruses. *J. Mol. Evol.*, 2008.

[48] Rafael Sanjuan et al. The distribution of fitness effects caused by single nucleotide substitutions in an rna virus. *PNAS*, 101 :8396–8401, 2004.

[49] Pedro F. Vale et al. The distribution of mutational fitness effects of phage ϕx174 on different hosts. *Evolution*, 2012.

[50] Thomas Bataillon Rees Kassen. Distribution of fitness effects among beneficial mutations before selection in experimental populations if bacteria. *Nature genetics*, 2006.

[51] Darin R. Rokyta et al. An empirical test of the mutational landscape model of adaptation using a single-stranded dna virus. *Nature genetics*, 37 :441–444, 2005.

[52] Paul Joyce et al. A general extreme value theory model for the adaptationn of dna sequences under strong selection and weak mutation. *Genetics*, 180 :1627–1643, 2008.

[53] Richard N. MacLaughlin. The spatial architecture of protein function and adaptation. *Nature*, 491 :138–142, 2012.

[54] Stuart Coles. *An introduction to statistical modeling of extreme values*. Springer, London, 2001.

[55] William Bialek Thierry Mora, Aleksandra M. Walczak and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA*, 107 :5405–5410, 2010.

[56] Eleni Katifori David Jordan, Seppe Kuehn and Stanislas Leibler. Behavioral diversity in microbes and low dimensional phenotypic spaces. *Proc. Natl. Acad. Sci. USA*, Early edition :1–6, 2013.

[57] Derek J. Smith et al. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305 :371–375, 2004.

[58] Alan Lapedes and Robert Farber. The geometry of shape space : application to influenza. *Santa Fe Institute*.

[59] Roonen Segev Elad Schneidman, Michael J. Berry and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440 :1007–1012, 2006.

[60] Mehran Kardar Jhon P. Barton and Arup K. Chakraborty. Scaling laws describe memories of host-pathogen riposte in the hiv population. *Proc. Natl. Acad. Sci. USA*, 112 :1965–1970, 2014.

Hierarchy and Extremes in Selections from Pools of Randomized Proteins

Sébastien Boyer,[1] Dipanwita Biswas,[1] Ananda Kumar Soshee,[1]
Natale Scaramozzino,[1] Clément Nizak,[2] and Olivier Rivoire[1]

[1]*Laboratoire Interdisciplinaire de Physique, CNRS & Université Grenoble Alpes, Grenoble, France*
[2]*Laboratoire de Biochimie, CNRS & ESPCI-ParisTech, Paris, France*

Variation and selection are the core principles of Darwinian evolution, yet quantitatively relating the diversity of a population to its capacity to respond to selection is challenging. Here, we examine this problem at a molecular level in the context of populations of partially randomized proteins selected for binding to well-defined targets. We built several minimal protein libraries, screened them *in vitro* by phage display and analyzed their response to selection by high-throughput sequencing. A statistical analysis of the results reveals two main findings: first, libraries with same sequence diversity but built around different "frameworks" typically have vastly different responses; second, the distribution of responses within a library follows a simple scaling law. We show how an elementary probabilistic model based on extreme value theory rationalizes these findings. Our results have implications for designing synthetic protein libraries, for estimating the density of functional biomolecules in sequence space, for characterizing diversity in natural populations and for experimentally investigating the concept of evolvability, or potential for future evolution.

Diversity is the fuel of evolution by natural selection but translating this concept into quantitative measurements is not straightforward [1]. A simple count of the number of different individuals in a population for instance fails to account for the very different responses to selection that two populations with same number of different individuals may elicit. The problem is already acute at the molecular scale where it also takes a very practical form: libraries of diverse proteins are routinely screened as a way to identify biomolecules of interest (binders, catalysts,...) and a proper "diversity" is critical for success [2, 3]. But beyond a general agreement that maximizing the number of different elements is desirable, there is no general rule for engineering and comparing diversity in these libraries.

A common design of many protein libraries is to concentrate variations at one or a few variable parts located around a fixed "framework", which is shared by all members of the library [2, 3]. The natural design of antibody repertoires, the pools of immune proteins with potential to recognize nearly every molecular target, follows this pattern. Most of sequence variations in antibodies are indeed concentrated at a few loops extending from a common structural scaffold [4]. This design has inspired the conception of artificial protein libraries built on frameworks other than the immunoglobulin fold [5].

Here, we propose a new approach to quantitatively characterize the selective potential of molecular libraries. To develop this approach, we designed and screened 24 synthetic protein libraries with identical sequence variations but different frameworks and analyzed their response to well-defined selective pressures by high-throughput sequencing. Between libraries, we find that selective potentials vary widely and define a hierarchy of frameworks. Within libraries, we find that selective potentials exhibit a simple scaling law, characterized by few parameters. The essence of these results is captured by an elementary probabilistic model based on extreme value theory (EVT). This leads us to propose a new measure of the selective potential of a population that overcomes the shortcomings of previously proposed measures of diversity.

I. METHODS

A. Library design

We built 24 minimal libraries with different frameworks but identical sequence diversity (Materials & Methods, Figures 1 and S1). Twenty frameworks consist of single-domain antibodies taken from natural heavy-chain genes of diverse origins (V_H fragments), typically sharing 40% of their amino acids (Figure S2); they originate from maturated antibodies, which are mutated relative to their germline form, except for the S1 framework, which comes from a germline (naïve) antibody. Three additional frameworks are more closely related and correspond to the germline and two maturated forms of the same human antibody, with the maturated frameworks sharing 65% and 85% sequence identity with the germline. Finally, one framework consists exclusively of glycines to serve as a control. Diversity is limited to four consecutive amino acids at the complementarity determining region 3 (CDR3), the part of antibody sequences most critical for specificity [6]. Structurally, the CDR3 forms one of three loops that define the binding pocket of a V_H domain [4]; in our design, the two other loops (CDR1 and 2) are thus part of the framework. Our libraries are minimal on two accounts: the framework consists of a single domain of ~ 100 amino acids and the total diversity is $20^4 = 1.6 \times 10^5$ – all combinations of the 20 natural amino acids at the 4 varied sites. For comparison, the most commonly used antibody libraries consist of two domains (V_H and V_L) and have $> 10^8$ variants, with variation introduced at different CDRs [7]. Libraries based on V_H only are, however, known to be effective [8].

"Minimalist libraries" have also been built by restricting the alphabet of amino acids at the variables sites but contained $> 10^{10}$ variants [9–11]. One of the simplest libraries demonstrated so far, built on a synthetic scaffold, still contained $> 10^6$ variants randomly sampled from a much larger pool of potential sequences [12].

B. Selection

We screened our libraries by phage display for binding to one of two targets, a neutral synthetic polymer, poly-vinylpyrrolidone (PVP), and a short DNA loop of 9 nucleotides (Materials & Methods). Two previous studies established the capacity of antibody phage display to select binders for these targets [2, 14]. Phage display is a standard high-throughput screening technique [15]. It is based on the fusion of each antibody sequence to the sequence of the pIII surface protein of the filamentous bacteriophage M13, a natural virus of the bacterium *E. coli* with the shape of a 1 μm long and 10 nm wide cylinder [15]. The engineered phage encapsulates the DNA sequence of an antibody and displays the corresponding polypeptide at its surface. Populations of up to 10^{14} phages displaying a total diversity of up to 10^{10} different antibodies can thus be manipulated. A round of selection consists in retrieving the phages bound either to the bottom of a plate where the PVP target is attached or to magnetic beads where the DNA target is coated. It is followed by a round of amplification achieved by infecting bacteria with the selected phages. We performed experiments where each sequence is initially present in $\sim 10^4$ copies and where targets are provided in excess. Starting either from a single library (single framework) or from a mixture of different libraries, three rounds of selection/amplification were performed. Although the enrichment of some of the sequences is intended to reflect binding to the specified targets, other factors may contribute, such as sequence-specific differences in amplification. In our experiments, such non target-specific selective factors can be detected but are non-dominant (Supporting Information). Our analysis and its interpretation, however, do not rely on the precise nature of the selective pressure.

C. High-throughput sequencing

We sequenced samples of $10^6 - 10^7$ sequences at different rounds of selection by Illumina Miseq pair-ended high-throughput sequencing [16]. The results give us an estimation of the relative frequencies f_i^t of each sequence i in the population at each round $t = 0, 1, 2, 3$. In estimating these frequencies, we take into account both sequencing and sampling errors (Materials & Methods).

We define the selectivity to a target of each sequence i as its probability s_i to pass a round of selection. This selectivity is inferred up to a multiplicative factor a as

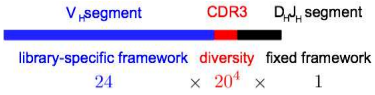


FIG. 1: Library design – We designed a total of 24 libraries with distinct frameworks and identical sequence diversity consisting of all $20^4 = 1.6 \times 10^5$ combinations of the 20 natural amino acids at 4 consecutive positions. The design follows the natural design of the variable (V) region of the heavy (H) chain of antibodies, which is assembled by joining three gene segments, the variable (V_H), diversity (D_H) and joining (J_H) segments. The library-specific parts of the frameworks (in blue) are from natural V_H and diversity is introduced at the third complementarity determining region (CDR3, in red), at the junction between V_H and $D_H J_H$, a part of the sequence critical for specific binding to antigens; the D_H and J_H segments (in black) are common to all libraries.

the ratio of the frequency of the sequence before and after selection [17]:

$$s_i = a \, \frac{f_i^t}{f_i^{t-1}}. \qquad (1)$$

The unknown multiplicative constant a reflects our lack of quantitative control over the rate of amplification of the sequences. In our calculations, we arbitrarily fix a so that $\sum_i s_i = 1$; we explain below how our conclusions depend on this choice. We compare the frequencies between rounds $t = 3$ and $t - 1 = 2$, where sequences with highest selectivities are best represented.

Previous studies have applied next-generation sequencing to the outcome of phage display screens as a way to identify a large number of binders [18, 19] but have not investigated the distribution of the relative selectivities of these binders.

D. Reproducibility and specificity

Several observations based on the frequencies and amino-acid patterns of the sequences in populations under selection validate our experimental approach. (1) Screening the same library against the same target in separate experiments yields reproducible frequencies f_i^t at the last round $t = 3$ (Figure S3). (2) Screening the same library against different targets yields target-specific amino-acid patterns (Figures S4). (3) Screening two libraries against the same target yields library-specific amino-acid patterns (Figure S4). Taken together, these results show that enrichment of some of the sequences is reproducible and arises from selection for specific binding to the targets. We note that one feature of our experiments is critical for reproducibility: the initial populations maximize degeneracy (the number of copies of each sequence) rather than diversity (the number of distinct sequences).

II. RESULTS

A. Hierarchy between libraries

To compare the selective potentials of libraries built around different frameworks, we performed experiments in which the initial population of sequences consists of a mixture of libraries with distinct frameworks – a meta-library. The results of these experiments reveal a striking hierarchy. Diverse members of a same library, i.e., sequences sharing a common framework, typically dominate. When repeating the experiment with an initial mixture of libraries that excludes the dominating library, another library dominates (Figure 2). Libraries not selected when mixed with other libraries nevertheless do contain sequences with detectable selectivities, as shown by screening them in isolation (Figure S4). These results are not explained by uneven representations of the libraries in the initial population or by framework-specific differences during amplification (Figure S5).

Differences in frameworks are thus generally more significant than differences between variable parts, even though these parts are clearly under selection for binding (Figures 3B and 3D). This result may not be surprising for very dissimilar frameworks, but our frameworks are all expected to share the same structural fold and some frameworks have few sequence differences. In particular, the dominating framework when selecting the mixture of all 24 libraries against the DNA target (Figure 2) is a germline human V_H framework, which dominates two libraries built on frameworks derived from it by affinity maturations, which share respectively 65% and 85% of their amino acids. The observed hierarchy is target dependent: different frameworks dominate when screening the meta-library against different targets. Remarkably, when screening the 24 libraries against the PVP target (Figure S6), the dominating framework is the only other germline framework of the mixture (the S1 framework). As noted previously, differences between frameworks also appear in the patterns of amino acids that are selected at the level of CDR3s (Figures S4C-E).

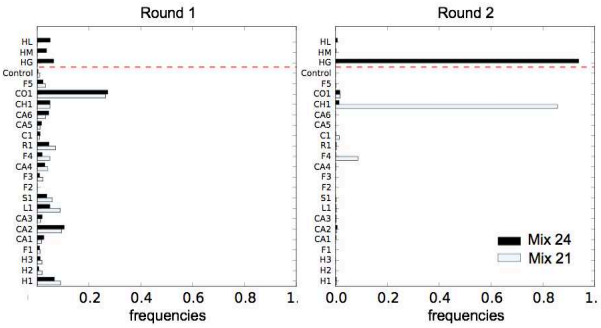


FIG. 2: Hierarchy between libraries – Frequencies of the different libraries in two successive rounds of selection against the DNA target. Black bars report selection of all 24 libraries and white bars selection of a subset of 21 libraries, excluding the 3 libraries above the red dotted line. At the second round (right), the population is enriched in sequences from one particular library, the HG library, in contrast to what is observed at the first round (left). The subset of 21 libraries excludes the library dominating the mixture of all 24 libraries, which leads another library, the CH1 library, to dominate. Within the two libraries, a diversity of CDR3 are selected (Figures 3B and 3D), with different patterns of amino acids (Figure S4). Enrichment from the other libraries can also be observed when they are screened in isolation (Supporting Information).

B. Scaling within libraries

To compare the selectivities of sequences sharing a common framework, and therefore differing by at most 4 amino acids (Figure 1), we rank these sequences in decreasing order of their selectivity s_i and plot these selectivities versus the ranks on a double logarithmic scale – a representation of the cumulative distribution of selectivities within a library. For several experiments, this representation reveals a power law: if $s(r)$ is the selectivity of the sequence of rank r, then, for the sequences with top ranks,

$$s(r) \sim r^{-\kappa}. \qquad (2)$$

Figure 3A shows an example where the exponent is $\kappa \simeq 0.5$. While this power law is observed for several libraries (different frameworks) and selective pressures (different targets), it is not systematic: deviations are often observed for the very top sequences (Figure 3B) and, for several experiments, a power law cannot be justified (Figure 3D).

Both the power law and its various deviations can, however, be rationalized under an elementary mathematical model. This model rests on two assumptions. First, it assumes that the selectivity of each sequence in a library is drawn independently at random from a common probability density $\rho(s)$, which may depend on the framework and the target. Second, it assumes that the sequences with top selectivities are in the tail of this probability density.

The model is thus probabilistic even though – baring out experimental noise – the experiments have no inherent stochastic element. To the extent that selectivity reflects binding at thermodynamical equilibrium, the selectivity s_i of antibody i is indeed determined by its binding free energy ΔG_i to the target: $s_i \propto e^{-\Delta G_i/k_B T}$, where T represents the temperature and k_B the Boltzmann constant. The binding free energy ΔG_i is a physical quantity which, in principle, is fully determined by the sequence of amino acids. In the spirit of applications of random matrix theory to nuclear physics [20], it may nevertheless be advantageous to discard this microscopic description in favor of a coarser probabilistic description, which treats the selectivities s_i as instances of random variables independently drawn from a common probabil-

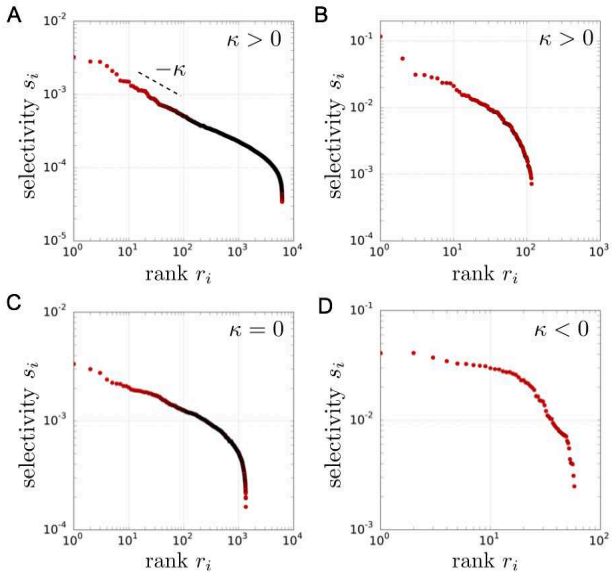


FIG. 3: Scaling relations within libraries – The selectivities s_i of the sequences are represented versus their ranks r_i for four experiences differing by the input library and the choice of the target against which it is selected. **A.** S1 library against the PVP target. **B.** HG library against the DNA target. **C.** F3 library against the PVP target. **D.** CH1 library against the DNA target. In A, the distribution of the top ~ 1000 sequences follows a power law with exponent $\kappa \simeq 0.5$. This behavior is consistent with the prediction of extreme value theory (EVT) when the shape parameter is positive, $\kappa > 0$ (see Figure 4 for the analysis that justifies this conclusion). Although not obvious from this representation, the data in B is also consistent with EVT when $\kappa > 0$ while the data in C and D are consistent with EVT when, respectively, $\kappa = 0$ and $\kappa < 0$.

ity density $\rho(s)$. In contrast to nuclear physics, no symmetry constrains $\rho(s)$ *a priori* but, if concerned only with the largest s_i, results from extreme value theory (EVT), the branch of probability theory dealing with extrema of random variables [21], do constrain the form of the tail of $\rho(s)$ from which they originate, thus allowing for non-trivial predictions.

EVT indeed indicates that random variables s independently drawn from the tail of a common probability density have themselves a probability density of the form [3]

$$f_{\kappa,\tau,s^*}(s) = f_\kappa\left(\frac{s-s^*}{\tau}\right), \tag{3}$$

with f_κ necessarily belonging to the generalized Pareto family:

$$f_\kappa(x) = \begin{cases} (1+\kappa x)^{-\frac{\kappa+1}{\kappa}} & \text{if } \kappa \neq 0, \\ e^{-x} & \text{if } \kappa = 0, \end{cases} \tag{4}$$

where the exponential for $\kappa = 0$ is just the continuous limit of $f_\kappa(x)$ when $\kappa \to 0$. Here, s^* represents a threshold above which the tail of $\rho(s)$ is defined, τ is a scaling factor (which absorbs the undetermined factor a introduced in Eq. (1)) and $\kappa \geq -1$ is the so-called shape factor (independent of a), which defines the universality class to which the distribution of selectivities belongs: the probability densities $\rho(s)$ may differ, but if they are associated with the same κ, events drawn form their tails will share similar statistical properties.

As suggested by the notations, when $\kappa > 0$, but only when $\kappa > 0$, this model predicts that the top ranked sequences follow a power law with exponent κ, as described by Eq. (8). Mathematically, when considering a large number N of samples, the rank $r(s)$ is indeed related to the cumulative distribution of selectivities by

$$r(s) \sim N \int_s^\infty \rho(x)\,\mathrm{d}x. \tag{5}$$

If $\rho(s) \sim s^{-\frac{\kappa+1}{\kappa}}$ for large s as predicted by Eq. (4) for $\kappa > 0$, we must then have $\int_s^\infty \rho(x)\,\mathrm{d}x \sim s^{-1/\kappa}$ and therefore $r(s) \sim s^{-1/\kappa}$, which is equivalent to Eq. (8). In other words, the power law seen in Figure 3A corresponds to the expected relationship between the rank and the values of random variables drawn from the tail of a probability density when this density belongs to a class associated with $\kappa > 0$.

To precisely assess the ability of our model to describe all these different cases, we followed the point-over-threshold approach, a standard method in applications of EVT to empirical data [3]. This approach consists in fitting the data s_i satisfying $s_i > s^*$ by a function of the form $f_{\kappa,\tau,s^*}(s)$ for different values of the threshold s^*, and then in estimating whether a threshold $s^*_{\min}$ exists such that for $s^* > s^*_{\min}$ the inferred parameter $\hat{\kappa}(s^*)$ is nearly independent of s^*. To apply this method, we inferred the parameters $\hat{\kappa}(s^*)$ and $\hat{\tau}(s^*)$ by maximum likelihood from the data $s_i > s^*$ for every value of s^*. For the data presented in Figure 3A, an illustration is provided in Figure 4A, with error-bars indicating 95% confidence intervals (see Supporting Information for the analysis of other experiments). In this example, we observe that $\hat{\kappa}(s^*)$ becomes nearly constant, of the order of 0.5, for $s^* > s^*_{\min} \simeq 4 \times 10^{-4}$. The determination of $s^*_{\min}$ is performed by visual inspection but any choice of $s^* > s^*_{\min}$ should give equivalent results.

Given $s^* > s^*_{\min}$ and the associated values of $\kappa = \hat{\kappa}(s^*)$ and $\tau = \hat{\tau}(s^*)$ inferred from maximum likelihood, the next step is to estimate whether this best fit is indeed a good fit. The diagnosis is commonly performed visually using probability-probability (P-P) and quantile-quantile (Q-Q) plots [3]. The P-P plot compares the empirical and modeled cumulative distributions by representing the quantile function $q(s) = r(s)/N$ (the fraction of the data above s) against the cumulative $F_{\hat{\kappa},\hat{\tau},s^*}(s) = \int_0^s f_{\hat{\kappa},\hat{\tau},s^*}(x)\,\mathrm{d}x$. As indicated by Eq. (5), a straight line $y = x$ is expected if the fit is perfect, which the inset of Figure 4B shows to be nearly the case in this example. The Q-Q plot makes a similar comparison but by representing s against $F^{-1}_{\hat{\kappa},0,0}(q^{-1}(s))$, where $q^{-1}(x)$ represents

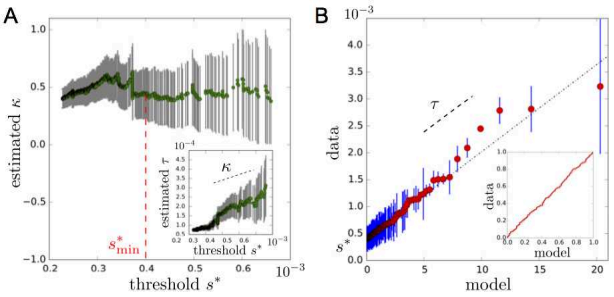


FIG. 4: Extreme value analysis by the point-over-threshold approach – **A.** Values of the inferred parameters $\hat{\kappa}(s^*)$ from selectivities $s_i > s^*$ as a function of the threshold s^*. The inference is made by maximum likelihood and the error-bars indicate 95% confidence intervals. Inset: similarly for $\hat{\tau}(s^*)$, the second parameter of the model, which is estimated jointly to $\kappa(s^*)$. For s^* sufficiently large, $s^* > s^*_{\min}$, $\hat{\kappa}(s^*)$ should be constant and $\hat{\tau}(s^*)$ increase linearly with slope $\hat{\kappa}(s^*)$. This is observed here for $s^*_{\min} \simeq 4 \times 10^{-4}$ (red dotted line) and leads to $\kappa = 0.45 \pm 0.22$ and $\tau = 1.6 \times 10^{-4} \pm 10^{-5}$; $\kappa = 0$ can be excluded by likelihood ratio test with a p-value $< 10^{-4}$. **B.** Quantile-quantile plot representing the data s_i against predictions from the model based on the inferred value of κ only. A straight line is expected for a good fit with a slope and the y-intercept given by the two other parameters τ and s^*. Inset: probability-probability plot comparing the empirical cumulative distributions from the data to the cumulative distribution from the inferred model, showing an excellent agreement. The data for this figure comes from the selection of the S1 library against the PVP target as in Figure 3A (see Figures S8-S10 for a similar analysis of the data shown in Figures 3B-D).

the value of s above which a fraction x of the data is located. This representation has two advantages over the P-P plot: it relies only on the estimation of κ and it displays more clearly the contribution of the most extreme values. A straight line is expected if the fit is perfect, but this time with a slope τ and a y-intercept s^*. The main panel of Figure 4B indicates again a very good fit in the illustrated case.

Performing the same analysis on results of selections of various libraries against various targets, we find that the model is able to describe all the experiments (Figures S8-S12). Different values of κ are obtained with differences that are statistically significant (Table S1). In particular, the three cases $\kappa > 0$, $\kappa = 0$ and $\kappa < 0$ are represented.

While many models can lead to a power law [23], our probabilistic model has the merit of explaining the various deviations from this behavior that the data exhibits. First, when $\kappa > 0$, EVT predicts a power law with exponent κ for the top-sequences but accounts for deviations both for the very top-ranked sequences, which, under the model, may vary widely (Figure S7), and for sequences of smaller selectivities, where f_κ in Eq. (4) can provide an excellent fit well beyond the point where the power law applies (e.g., Figures 4B and S8). Second, EVT predicts behaviors differing from a power law if the probability density $\rho(s)$ belongs to an universality class associated with $\kappa \leq 0$, consistently with the results of some of the experiments (e.g., Figures 3D and S10).

III. DISCUSSION

We presented a quantitative analysis of *in vitro* selections of multiple libraries of partially randomized proteins with variations limited to four consecutive amino acids. The distribution of selectivities of the top sequences is described by few parameters, with an interpretation provided by an elementary probabilistic model based on extreme value theory (EVT).

Within a library whose members share a common framework, this distribution is characterized by a shape parameter κ, which may be either positive, negative or zero. This parameter is independent of the unknown factor a in Eq. (1) and has several interpretations. For instance, it controls the relative spacing between selectivities: ranking the sequences from best to worst, the expected difference of selectivity between sequences at rank r and $r+1$, $\Delta_r = \mathbb{E}[s_r - s_{r+1}]$, satisfies $\Delta_r/\Delta_1 \sim r^{-(\kappa+1)}$, i.e., the larger κ, the wider the spread between phenotypes in the library (Supporting Information). The shape parameter also provides a statistical answer to the following question: if sampling N sequences yields a top sequence of selectivity s_1, what best selectivity s'_1 may we expect from sampling $N' > N$ sequences? The difference $\mathbb{E}[s'_1 - s_1]$ is a sharply increasing function of κ (Supporting Information; Figure S13); as consequence, multiplying by a factor 1000 the number of sequences when $\kappa = 0$ is expected to have same effect as multiplying it by a factor 2 when $\kappa = 0.2$, if starting with $N = 10^5$ sequences.

Besides the shape parameter κ, the other parameters are the scaling parameter τ, the threshold of selectivities s^* that defines where the tails starts, and the fraction ϕ of the data above this threshold (there is some freedom in the choice of s^* on which both τ and ϕ depend: see Supporting Information). Within our experimental set-up where the selectivities are determined only up to a multiplicative factor (see Eq. (1)), the values of s^*, ϕ and τ obtained from different experiments cannot be directly compared, but our selections with mixtures of libraries suggest that s^* varies from library to library on a scale larger than the scale of the differences of selectivity within libraries. All the parameters of the model are found to be both framework and target dependent (Table S1).

Based on these results, we propose these parameters as general descriptors of the selective potential of a population of random variants facing a given selective constraint. In particular, these descriptors could be applied to re-visit the fundamental problem of estimating the density of functional proteins or RNAs in sequence space. Previous studies have estimated this density by counting the number of different sequences enriched in *in vitro* selections [24, 25]. The results of such experiments depend

on experimental noise, which sets a lower limit s_{noise} on detectable selectivities. In turn, our approach is dependent only on the library content and the selective pressure, provided $s^* > s_{\text{noise}}$.

Power laws are seemingly ubiquitous in distributions of protein features [26, 27]. Most closely related to our work, the distribution of abundances of distinct antibody sequences in Zebrafishes has been shown to follow a power law with exponent $\alpha \simeq 1$ [28, 29]. Only instantaneous frequencies, not selectivities, are accessible in such a case, but, assuming a homogeneous initial distribution of sequences, frequencies and selectivities have same distribution and $\alpha = \kappa$ if $\kappa > 0$. However, repeating n times the same selection leads to $\alpha = n\kappa$, which does not account for a stable exponent $\alpha > 0$, which may arise in natural repertoires from fluctuating selective pressures [30]. One possible extension of our approach could be to explore this scenario by changing the target between successive rounds of selection.

While many models can be consistent with a power law, our model based on EVT covers without additional assumption the deviations from a power law observed in the data; in particular, it can fit the data over a larger range of selectivities and account for non-power-law behaviors. This is, however, not the first application of EVT to the description of biological variation: Gillespie first introduced it in models of evolutionary dynamics as a way to constrain the distribution of beneficial effects obtained when mutating a wild-type individual [31, 32]. He assumed $\kappa = 0$, arguing that this class includes all "well-behaved" distributions, among which the exponential, normal, log-normal and gamma distributions. Mathematical models for the distribution of affinities in combinatorial molecular libraries have also proposed that it should have universal features but only considered distributions in the exponential class $\kappa = 0$ [33, 34].

Several experimental studies have recently investigated the value of κ applicable to the distribution of beneficial effects in viral or bacterial populations [35, 36]. The sample sizes available in these studies are, however, insufficient to conclusively validate or invalidate the EVT hypothesis. In these experiments, the number of mutants found in the tail has indeed been so far very low, of the order of a dozen: estimating the sign of the shape parameter κ can be attempted [37], but assessing the validity of the fit using quantile-quantile plots as in Figure 4 is simply impossible with such limited data. Our rich dataset thus provides the first thorough test of the applicability of EVT to the analysis of biological diversity.

Comparable datasets are now being increasingly produced. In particular, several groups have characterized the phenotype of every single-point mutant of a protein [38]. Our model may be viewed as a mathematical formalization of the concept of a random library, from which single-points mutants may deviate. We note, however, that selectivities from non-random subsets of one of our libraries do follow the same model as the full library (Figure S14). In any case, significant deviations will have to be quantified against our null model.

Beyond protein libraries, the model is relevant to the screening of synthetic chemical libraries, including the combinatorial libraries of small molecules developed in the pharmaceutical industry for drug discovery [39, 40]. In this context, one previous study was performed with enough data points to possibly discriminate between different universality classes [41], although the authors considered only the exponential case $\kappa = 0$.

Finally, our work raises a new question for future studies: if the selective potential of a partially randomized library is captured by few parameters, and if these parameters can vary from library to library, what controls them? More simply, what features of the framework define a universality class? For instance, how extending the variable parts to other sites changes κ? The patterns of amino acids forming the sequences, which we have analyzed here only to confirm the reproducibility of the experimental results and their specificity with respect to the targets and libraries, may provide valuable insights [29].

The question may also be asked at another level: can we or natural evolution control these parameters to optimize the selective potential of a population? This question relates to the debated "evolution of evolvability" [42, 43], cast here into a concrete conceptual and experimental setting. Antibodies potentially define an excellent model system to study experimentally this question, since they are subject to selection and maturation towards a diversity of targets as part of their natural function. The approach and concepts introduced in this work provide the means to address the problem with quantitative experiments.

Materials & Methods

Phage Display – PVP-plates were prepared as described in [2]. The DNA target was prepared by self-assembly of a hairpin DNA, labeled with biotin at its 5' end (5'-Biotin– AAAAGACCCCATAGCGGTCTGCGT), and was purchased from Eurogentec (Angers, France). *E. coli* TG1 competent cells were purchased from Lucigen Lt. Phage production, phage-display screens based on the pIT2 phagemid vector and helper phage KO7 production were performed following the standard protocol from Source BioScience (Cambridge, U.K.; http://lifesciences. sourcebioscience.com/media/143421/tomlinsonij.pdf) and our own previous work [2, 44] with some modifications specified in the Supporting Information.

Sequencing data – Library phagemids were purified from E. coli stocks after each selection round using Midiprep kits from Macherey-Nagel (Hoerdt, France). v3 Illumina MiSeq sequencing was performed by Eurofins Genomics (Ebersberg, Germany). The MiSeq pair-ended technology was used. Frameworks were recovered on the forward read and only the reads having all the expected restriction sites and less than 4 errors on the 126 bases were kept. The CDR3 were accessible on the reverse read and only the reads having all

the expected restriction sites and a average value of quality read Q>30 on the 12 bases defining the CDR3 were kept (see Table S2 for an estimation of sequencing errors).

Computational analysis – We infer the selectivity s_i of an amino acid sequence i by Eq. (1) with $t = 3$ (third round of selection). The frequencies are simply given by $f_i^t = n_i^t / \sum_j n_j^t$ where n_i^t is the number of sequences i present in the sample. Given sampling errors, estimated as $\Delta s_i/s_i = 1/\sqrt{n_i^2} + 1/\sqrt{n_i^3}$, and given sequencing errors, estimated at $\sim 5\%$ over the 12 bases of the CDR3 (Table S2), the estimation of s_i is meaningful only for sequences that are sufficiently present at each round: $n_i^{t-1} > n_0$ and $n_i^t > n_0$. We took $n_0 = 10$ and verified that the results are not sensitive to this exact value (Table S3). With $n_0 = 10$, relative sampling errors are, in the worst case, as high as $2/\sqrt{n_0} \sim 60\%$, but, assuming that sampling errors are uncorrelated, this uncertainty has no major incidence on the estimation of aggregated properties of the distribution of the largest s_i, which involves several hundreds of different i.

Extreme value statistics – We followed the standard approach for modeling threshold excesses [3]. The parameters κ and τ were estimated by maximum likelihood and the 95% confidence intervals shown in Figure 4A were obtained under the hypothesis of normality by calculating the inverse of Fisher's information. To ensure that the data allows us to discriminate between $\kappa = 0$ and $\kappa \neq 0$, a p-value was calculated by a likelihood ratio test, whose distribution was estimated both by numerical simulations. Maximum likelihood estimations are calculated on at least 50 data points.

Author contributions

SB, CN and OR designed research; AKS set up phage display; AKS and OR designed the libraries; DB constructed the libraries; SB performed the selections; CN and NS supervised the experiments; SB and OR analyzed data; SB, CN and OR wrote the paper. DB, AKS and NS contributed equally to this work.

Acknowledgments

This work was supported by Agence Nationale de la Recherche (ANR-10-PDOC-004-01, to O.R.) and by AXA Research Fund (post-doctoral grant to D.B). We thank S. Girard, B. Houchmandzadeh, T. Mora, R. Ranganathan and A. Walczak for helpful discussions.

[1] A E Magurran. *Measuring biological diversity*. John Wiley & Sons, 2013.

[2] H Zhao and F H Arnold. Combinatorial protein design: strategies for screening protein libraries. *Current Opinion in Structural Biology*, 7(4):480–485, 1997.

[3] T S Wong, D Zhurina, and U Schwaneberg. The diversity challenge in directed protein evolution. *Combinatorial chemistry & high throughput screening*, 9(4):271–288, 2006.

[4] E A Padlan. Anatomy of the antibody molecule. *Mol Immunol*, 31(3):169–217, 1994.

[5] A Urvoas, M Valerio-Lepiniec, and P Minard. Artificial proteins from combinatorial approaches. *Trends Biotechnol*, 30(10):512–520, 2012.

[6] J L Xu and M M Davis. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity*, 13(1):37–45, 2000.

[7] H R Hoogenboom. Selecting and screening recombinant antibody libraries. *Nature Biotechnology*, 23(9):1105–1116, 2005.

[8] E S Ward, D Güssow, A D Griffiths, P T Jones, and G Winter. Binding activities of a repertoire of single immunoglobulin variable domains secreted from Escherichia coli. *Nature*, 341(6242):544–546, 1989.

[9] F A Fellouse, C Wiesmann, and S S Sidhu. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc Natl Acad Sci USA*, 101(34):12467–12472, 2004.

[10] F A Fellouse, B Li, D M Compaan, A A Peden, S G Hymowitz, and S S Sidhu. Molecular recognition by a binary code. *J Mol Biol*, 348(5):1153–1162, 2005.

[11] F A Fellouse, K Esaki, S Birtalan, D Raptis, V J Cancasci, A Koide, P Jhurani, M Vasser, C Wiesmann, A A Kossiakoff, S Koide, and S S Sidhu. High-throughput Generation of Synthetic Antibodies from Highly Functional Minimalist Phage-displayed Libraries. *J Mol Biol*, 373(4):924–940, 2007.

[12] M A Fisher, K L McKinley, L H Bradley, S R Viola, and M H Hecht. De novo designed proteins from a library of artificial sequences function in Escherichia coli and enable cell growth. *PLoS ONE*, 6(1):e15364, 2011.

[2] A K Soshee, S Zürcher, N D Spencer, A Halperin, and C Nizak. General In Vitro Method to Analyze the Interactions of Synthetic Polymers with Human Antibody Repertoires. *Biomacromolecules*, 15(1):113–121, 2014.

[14] S Modi, C Nizak, S Surana, S Halder, and Y Krishnan. Two DNA nanomachines map pH changesalong intersecting endocytic pathways insidethe same cell. *Nature Nanotech*, 8(6):459–467, 2013.

[15] G Smith and V Petrenko. Phage Display. *Chem Rev*, 97:391–410, 1997.

[16] J Shendure and H Ji. Next-generation DNA sequencing. *Nature Biotech*, 26(10):1135–1145, 2008.

[17] D M Fowler, C L Araya, S J Fleishman, E H Kellogg, J J Stephany, D Baker, and S Fields. High-resolution mapping of protein sequence-function relationships. *Nature Methods*, 7(9):741–746, 2010.

[18] E Dias-Neto, D N Nunes, R J Giordano, J Sun, G H Botz, K Yang, J C Setubal, R Pasqualini, and W Arap. Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis. *PloS one*, 4(12):e8338, 2009.

[19] U Ravn, F Gueneau, L Baerlocher, M Osteras, M Desmurs, P Malinge, G Magistrelli, L Farinelli, M H Kosco-Vilbois, and N Fischer. By-passing in vitro screening–next generation sequencing technologies applied

to antibody display and in silico candidate selection. *Nucleic Acids Res*, 38(21):e193, 2010.
[20] ML Mehta. *Random matrices and the statistical theory of energy levels*. Academic press, 1967.
[21] E J Gümbel. *Statistics of extremes*. Columbia Univ. Press, 1958.
[3] S Coles. *An introduction to statistical modeling of extreme values*. Springer, 2001.
[23] M Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 2(1):226–251, 2004.
[24] A D Ellington and J W Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6287):818–822, 1990.
[25] A D Keefe and J W Szostak. Functional proteins from a random-sequence library. *Nature*, 410(6829):715–718, 2001.
[26] M A Huynen and E van Nimwegen. The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol*, 15(5):583–589, 1998.
[27] E V Koonin, Y I Wolf, and G P Karev. The structure of the protein universe and genome evolution. *Nature*, 420(6912):218–223, 2002.
[28] J A Weinstein, N Jiang, R A White, D S Fisher, and S R Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–810, 2009.
[29] T Mora, A M Walczak, W Bialek, and C G Callan. Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA*, 107(12):5405–5410, 2010.
[30] J Desponds, T Mora, and A M Walczak. Fluctuating fitness shapes the clone size distribution of immune repertoires. *arXiv preprint arXiv:1507.00751*, 2015.
[31] J H Gillespie. A randomized sas-cff model of natural selection in a random environment. *Theor Pop Biol*, 21(2):219–237, 1982.
[32] J H Gillespie. *The causes of molecular evolution*. Oxford University Press, 1991.
[33] Do Lancet, E Sadovsky, and E Seidemann. Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system. *Proc Natl Acad Sci USA*, 90(8):3715–3719, 1993.
[34] M M Tanaka, S A Sisson, and G C King. High affinity extremes in combinatorial libraries and repertoires. *J Theor Biol*, 261(2):260–265, 2009.
[35] C Beisel, D Rokyta, H Wichman, and P Joyce. Testing the extreme value domain of attraction for distributions of beneficial fitness effects. *Genetics*, 176(4):2441–2449, 2007.
[36] T Bataillon and S F Bailey. Effects of new mutations on fitness: insights from models and data. *Ann N Y Acad Sci*, 1320(1):76–92, 2014.
[37] D R Rokyta, C J Beisel, P Joyce, M T Ferris, C L Burch, and H A Wichman. Beneficial Fitness Effects Are Not Exponential for Two Viruses. *J Mol Evol*, 67(4):368–376, 2008.
[38] D M Fowler and S Fields. Deep mutational scanning: a new style of protein science. *Nat Methods*, 11(8):801–807, 2014.
[39] S L Schreiber. Organic chemistry: Molecular diversity by design. *Nature*, 457(7226):153–154, 2009.
[40] W R J D Galloway, A Isidro-Llobet, and D R Spring. Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nat Commun*, 1(6):1–13, 2010.
[41] S S Young, C F Sheffield, and M Farmen. Optimum utilization of a compound collection or chemical library for drug discovery. *J Chem Inf Model*, 37(5):892–899, 1997.
[42] G P Wagner and L Altenberg. Complex adaptations and the evolution of evolvability. *Evolution*, 50:967–976, 1996.
[43] M Pigliucci. Is evolvability evolvable? *Nat Rev Gen*, 9(1):75–82, 2008.
[44] P Jain, A Soshee, S S Narayanan, J Sharma, C Girard, E Dujardin, and C Nizak. Selection of Arginine-Rich Anti-Gold Antibodies Engineered for Plasmonic Colloid Self-Assembly. *J Phys Chem C*, 118:14502–14510, 2014.

SUPPORTING INFORMATION

Supplementary experimental methods

Library construction

The library-specific parts of the frameworks, upstream of the variable CDR3 (Figure 1) are shown in Figure S21. 23 of these frameworks were designed based on the amino acid sequences of 23 natural V_H segments, with minor modifications to accommodate common restriction sites at the two ends of the CDR2 and CDR3. Out of these 23 frameworks, 20 were chosen to have minimal sequence similarities, and 3 are from a same human V_H segment: one is the germline (naïve) form, one results from limited maturation (85% sequence similarity to the germline) and the other from extensive maturation (broadly neutralizing antibody against HIV [1] with only 65% sequence similarity to the germline). A 24th framework was made exclusively of glycines to serve as a control. Downstream of the CDR3, the fixed part has amino acid sequence FDYWGQGTLVTVSSG in all libraries. The nucleotide sequences were optimized for *E. coli* codon usage and are provided as supplementary file.

The 24 frameworks were obtained from Genewiz (South Plainfield, NJ) as synthetic genes with restriction sites flanking the CDRs to allow for the introduction of arbitrary sequences at the CDRs. In particular, the CDR3 region is flanked by BssHI and XhoI sites. These synthetic genes were cloned into a modified version of pIT2 phagemid (standard phage display vector) lacking V_L [2]. To randomize the CDR3 region, a degenerate oligonucleotide containing 12 random nucleotides (from Eurogentec, Angers, France) flanked by BssHI and XhoI sites was PCR-amplified, digested, and ligated into gel purified pIT2 phagemids harboring each of the 24 frameworks. Ligation products were purified and electroporated into TG1 *E. coli* (from Lucigen, Middleton, WI) at efficiencies exceeding 10^7 transformants (to ensure a >100-fold coverage of the 10^5 diversity), while keeping 100-fold lower efficiency in control electroporations of ligation product without insert (to minimize the occurence of empty vectors in libraries, below 1%).

Phage display screening

All chemicals were purchased from Sigma-Aldrich (St Louis, MO) unless otherwise specified. Deionized water of resistivity 16 MΩ.cm was produced with an ion exchange resin (Aquadem(R) system, Veolia, Lyon, France). 2xTY medium was prepared by dissolving 16 g tryptone, 10 g yeast extract, 5 g NaCl (tryptone and yeast extract from USBIO distributed by Euromedex, Strasbourg, France) in 1 L of deionized H2O and autoclaving for 15 min at 120 °C.

The DNA target (PAGE purified, lyophilized) was resuspended with deionized water at 400 μM. 20 mg of magnetic beads coated with streptavidin (Dynabeads(R) M-280 Streptavidin from Invitrogen Life Technologies SAS, Saint Aubin, France) were prepared according to the manufacturer's protocol. 10 μL of DNA stock solution were mixed with 20 mg of washed Dynabeads(R) and incubated for 10 minutes at room temperature using gentle rotation. The biotinylated hairpin DNA coated beads were separated with a strong magnet for 2-3 minutes and washed 2-3 times with a buffer containing 5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA and 1M NaCl.

Phage production is the same as described before except that the infected TG1 culture was grown for 7 hours (instead of overnight) at 30°C in 2xTY + 100 μg/mL ampicillin + 50 μg/mL kanamycin.

During phage display experiments, the supernatant containing our library (around 10^{11} phages) in 2xTY, was adjusted to 10 mM NaPO4 pH=7.4. Phages were first incubated against either naked magnetic beads or non-treated polystyrene 3 cm diameter Nunc Petri dish (Thermo Fisher Scientific, Waltham, MA) for negative selection. For DNA target selection, DNA LoBind tubes (Eppendorf AG, Hamburg, Germany) were used. Phages were incubated during 1 h without agitation and 30 min on a rocker at room temperature. The remaining phages were then incubated with either hairpin DNA or PVP targets. In the case of hairpin DNA, 50 μL of beads were incubated with an excess of DNA targets (around 10^{14}), washed according to the manufacturer's protocol, yielding on the order of 10^{13} immobilized DNA targets, at a 100-fold excess over available phages (10^{11}). Antibody selection was then performed against either DNA-coated beads or a PVP-functionalzed Petri dish for 90 mn on a rocker. 10 washing steps with 1xPBS + 0.1%Tween 20 were performed. Next, selected phages were eluted using 1 mL of fresh solution of 100 mM triethylamine for 20 min and neutralized with 500 μL of Tris/HCl buffer (1 M, pH 7.4). Eluted phages were rescued by infection of an excess of exponentially growing TG1 *E. coli* cells (14 mL of a 2xTY culture at O.D. 600 nm = 0.6)

for titration and phage preparation for subsequent rounds of selection. Infected TG1 were then plated on 2xTY + ampicillin plates for overnight amplification at 37°C. Glycerol stocks were stored at −80°C.

Amplification biases

Each round of selection is followed by a round of amplification consisting in infecting the bacteria with the selected phages. Sequence-specific differences in amplification may arise from differences of growth rate of the bacteria carrying different phagemids, or differences in infectivity or display ability of the phages. We measured by sequencing both the differences between frameworks when considering a mixture of the 24 libraries (Figure S9) and between CDR3 when considering a library of given framework (Figure S19).

Between frameworks, only the S1 and CH1 libraries, show significant enrichment upon amplification alone. Each of these two libraries dominates over the others in one experiment of selection with a mixture of libraries but, when the mixture of all 24 libraries is selecting against the PVP target, they are dominated by another library, the HG library, which does not show any enrichment upon amplification. This observation, together with the strong correlation between frequencies before and after amplification (Figure S9B), are evidence that differences in library amplification are not responsible for the observed hierarchy between frameworks (Figure 2).

Within each library, a clear enrichment for the glutamine amino acid is observed, irrespectively of the framework (Figure S19). This bias has a simple interpretation: the *supE* strain of *E. Coli* that we use for phage display is a partial amber stop codon suppressor. In this strain, the amber codon codes about one third of the times for a glutamine and acts as a stop codon the two other thirds. The reduced production of antibodies due to the presence of an amber codon thus confers a growth advantage to the bacteria (antibody expression is costly for *E. coli*). Consistently with this interpretation, we verify that all the glutamines present in the data are associated with the amber codon. The results presented in the paper exclude sequences with an amber codon but, in most experiments with selection, glutamine does not appear in the selected consensus sequence and considering the amber code as coding for an amino acid or for a stop codon has no incidence on the conclusions. Apart from glutamine amplification, no other significative pattern of amplification is visible or may plausibly explain the results of the experiments with selection.

Supplementary properties of extreme value distributions

Relations between parameters

The fit to an extreme value distribution with parameters (κ_0, τ_0) applies for selectivities above a threshold s_0^*. Fitting the data above a larger threshold $s_1^* > s_0^*$ must lead to the same shape parameter $\kappa_1 = \kappa_0$ (simply denoted κ) but to a different scaling parameter τ_1 given by [3]

$$\tau_1 = \tau_0 + \kappa(s_1^* - s_0^*). \tag{S6}$$

These are the relations verified in Figure 4A.

Another independent parameter, which depends on the bulk of the distribution, is the fraction ϕ_0 of the data above the threshold s_0^*, which obviously depends on the value of the threshold ($\phi_1 < \phi_0$ when $s_1^* > s_0^*$).

In total, four parameters are thus relevant: s^*, ϕ, κ and τ.

Spacings between extremes

We show here that if $s_1 > s_2 > \dots$ are drawn at random with a probability density $f_{\kappa,\tau,s^*}(s)$ given by Eq. (3) then their spacings defined by $\Delta_r = s_r - s_{r+1}$ scale as $\Delta_r/\Delta_1 \sim r^{-(\kappa+1)}$, where κ is the shape parameter. This follows from a more general result:

$$\Delta_r \sim \tau N^\kappa r^{-(\kappa+1)}, \tag{S7}$$

where τ is the scaling parameter and N the total number of samples.

The proof can be given in terms of the rescaled variable $x = (s - s^*)/\tau$ whose probability density $f_\kappa(x)$ is defined in Eq. (4), since $\Delta_r = \tau(x_r - x_{r+1})$. As indicated by Eq. (5), the rank r and the value x are related for large N by $r(x)/N \sim \int_x^\infty f_\kappa(u)\mathrm{d}u = (1+\kappa x)^{-1/\kappa}$. Inverting this relation gives $x_r = q(r/N)$ with $q(z) = (1 - z^{-\kappa})/\kappa$. In the limit of large N where the formalism applies, we have therefore $x_r - x_{r+1} \simeq -(1/N)q'(r/N)$ with the derivative of $q(z)$ given by $q'(z) = -z^{-(\kappa+1)}$. All together, this gives $x_r - x_{r+1} \simeq N^\kappa r^{-(\kappa+1)}$ and thus $\Delta_r \sim \tau N^\kappa r^{-(\kappa+1)}$.

Scaling of maxima with library sizes

We show here that if the maximum of N random variables drawn with a probability density given by Eq. (3) is s_1, then adding more elements to produce a library $m > 1$ times larger leads to a maximal value $s_1' \geq s_1$ satisfying

$$\mathbb{E}[s_1' - s_1] = \tau N^\kappa m^\kappa \mathrm{Li}_{\kappa+1}\left(1 - \frac{1}{m}\right), \tag{S8}$$

where $\mathrm{Li}_k(z) = \sum_{n=1}^\infty z^n/n^k$ defines the so-called polylogarithmic function. $\mathbb{E}[s_1' - s_1]$ is thus an increasing function of κ as illustrated in Figure S17, where Eq. (S8) is also compared to numerical simulations. The relevance of this formula rests on the assumption that sub-libraries of a library with shape parameter κ are characterized by the same κ, which finds support in the data (Figure S18). In practice, the extreme value distribution applies only to the fraction ϕN of the data above the threshold s^*. When expressed relative to the expected spacing between the top two values in the initial population of N variables, $\Delta_1 = \tau N^\kappa$, Eq. (S8) is however independent of N and τ: $\mathbb{E}[s_1' - s_1]/\Delta_1 = m^\kappa \mathrm{Li}_{\kappa+1}\left(1 - \frac{1}{m}\right)$.

To derive this formula, we consider an initial population of size mN whose maximum is s_1' and define a subpopulation of (approximate) size N by retaining with probability $1/m$ each of its elements. The maximum s_1 of this subpopulation has thus rank n in the initial population with probability $p_n = (1-1/m)^{n-1}1/m$ – the probability that none of $n-1$ top values are retained but that the n-th is. Following Eq. (S7), the distance between $s_1 = s_n'$ and s_1' is estimated as $\delta_n' = s_1' - s_n' = \sum_{r=1}^{n-1} \Delta_r' = \tau(mN)^\kappa \sum_{r=1}^{n-1} r^{-(\kappa+1)}$. This leads to

$$\mathbb{E}[s_1' - s_1] = \sum_{n=1}^\infty p_n \delta_n' = \frac{\tau(mN)^\kappa}{m} \sum_{n=1}^\infty \sum_{r=1}^{n-1} \left(1 - \frac{1}{m}\right)^{n-1} \frac{1}{r^{\kappa+1}} = \frac{\tau(mN)^\kappa}{m} \sum_{r=1}^\infty \sum_{n=r+1}^\infty \left(1 - \frac{1}{m}\right)^{n-1} \frac{1}{r^{\kappa+1}} \tag{S9}$$

and, after summing the geometric series $\sum_{n=r+1}^\infty (1-1/m)^{n-1} = (1-1/m)^r m$, to

$$\mathbb{E}[s_1' - s_1] = \tau(mN)^\kappa \sum_{n=1}^\infty \left(1 - \frac{1}{m}\right)^{n-1} \frac{1}{r^{\kappa+1}}, \tag{S10}$$

which is equivalent to Eq. (S8).

[1] F Klein, R Diskin, J F Scheid, C Gaebler, H Mouquet, I S Georgiev, M Pancera, T Zhou, R-B Incesu, B Z Fu, P N P Gnanapragasam, T Y Oliveira, M S Seaman, P D Kwong, P J Bjorkman, and M C Nussenzweig. Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell*, 153(1):126–138, 2013.

[2] A K Soshee, S Zürcher, N D Spencer, A Halperin, and C Nizak. General in vitro method to analyze the interactions of synthetic polymers with human antibody repertoires. *Biomacromolecules*, 15(1):113–121, 2014.

[3] S Coles. *An introduction to statistical modeling of extreme values*. Springer, 2001.

Supplementary tables

	κ	τ	s^*	ϕ
S1/PVP	0.44 ± 0.22	$1.6 \times 10^{-4} \pm 10^{-5}$	0.37×10^{-3}	$\sim 6 \times 10^{-4}$
F3/PVP	0.07 ± 0.21	$3.1 \times 10^{-4} \pm 8 \times 10^{-5}$	1.2×10^{-3}	$\sim 6 \times 10^{-4}$
HG/DNA	0.26 ± 0.21	$5.7 \times 10^{-3} \pm 1.5 \times 10^{-3}$	0.7×10^{-3}	$\sim 6 \times 10^{-4}$
CH1/DNA	-0.62 ± 0.25	$2.5 \times 10^{-2} \pm 8 \times 10^{-3}$	2.5×10^{-3}	$\sim 3.6 \times 10^{-4}$

TABLE SI: Parameters κ, τ, s^* and ϕ describing the four experiments presented in Figure 3. Note that τ and ϕ depend on s^*, which may be chosen within a finite interval of values. However, the values of $\tau(s^*)$ and $\phi(s^*)$ at $s^* = s_0^*$ determine their values at $s^* = s_1^*$ as indicated in Eq. (S6) for τ.

	Fraction of sequences with >1 sequencing error / 12 bases
Mix24	0.043
Mix24 amplified	0.029
Mix24/PVP round 1	0.025
Mix24/PVP round 2	0.051
Mix24/PVP round 3	0.107
Mix24/DNA round 1	0.032
Mix24/DNA round 2	0.065
Mix24/DNA round 3	0.029
Duplicate Mix24/DNA round 3	0.024
Mix21/PVP round 2	4×10^{-3}
Mix21/PVP round 3	4×10^{-3}
Mix21/DNA round 1	0.027
Mix21/DNA round 2	0.046
Mix21/DNA round 3	0.106
F3/PVP round 1	0.034
F3/PVP round 2	0.029
F3/PVP round 3	0.04
F3/DNA round 1	0.027
F3/DNA round 2	0.048
F3/DNA round 3	0.085

TABLE SII: Estimation of sequencing errors – Fraction of the sequences with at least one error in the 12 bases immediately downstream of the 12 based of the CDR3 (errors estimated given the known sequence of the fixed part of the framework).

S1/PVP in Mix24 $n_0^2 = n_0^3 = 10$	$\kappa = 0.34 \pm 0.22$
S1/PVP in Mix24 $n_0^2 = 10$ and $n_0^3 = 25$	$\kappa = 0.42 \pm 0.25$
S1/PVP in Mix21 $n_0^2 = n_0^3 = 100$	$\kappa = 0.56 \pm 0.18$
S1/PVP in Mix21 sampled $n_0^2 = n_0^3 = 10$	$\kappa = 0.48 \pm 0.16$
S1/PVP in Mix21 sampled $n_0^2 = 10$ and $n_0^3 = 50$	$\kappa = 0.67 \pm 0.37$

TABLE SIII: Robustness of the EVT analysis – The analysis presented in the main text retains only sequences present in sufficient number in the samples of the populations that are sequenced at the second and third rounds – namely $n_i^2 > n_0^2 = 10$ and $n_i^3 > n_0^3 = 10$. This table shows that varying the values of the thresholds n_0^2 and n_0^3 has little incidence on the value of the shape parameter κ inferred by EVT analysis. The sample of the S1 library against PVP in the mixture of 21 libraries (Mix21, see Figure 2) contained 10^6 sequences while the sample in the mixture of 24 libraries (Mix24) contained only 10^5; the two last rows of the table shows that further sampling at 1/10 the former to reach samples of comparable sizes have no incidence on the results.

Supplementary figures

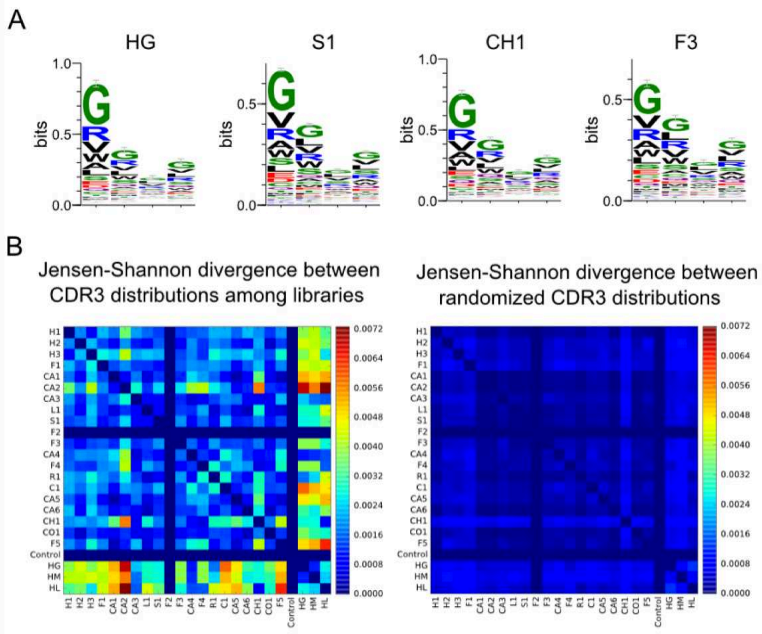


FIG. S5: Diversity of the libraries – The different libraries are intended to harbor the same distribution of amino acids at the 4 varied positions. We measured these distributions by sequencing samples from the initial libraries. **A.** Sequence logos showing the entropies of the various amino acids at the four positions: the distribution is non-uniform but similar from one library to the next. **B.** More quantitatively, the distance between distributions is estimated using the Jensen-Shannon divergence: if q_a^ℓ is the frequency of amino acid a in the CDR3 of library ℓ, the Jensen-Shannon divergence between libraries k and ℓ is defined as $\sum_a q_a^k \ln(q_a^k/q_a^\ell) + \sum_a q_a^\ell \ln(q_a^\ell/q_a^k)$. This divergence is found to be 5 to 10 times larger than expected from sampling noise. This represents the experimental precision at which we were able to introduce the same diversity in each library. These differences of frequencies between initial libraries are, however, much smaller than the differences of frequencies before and after a round of selection within a same library.

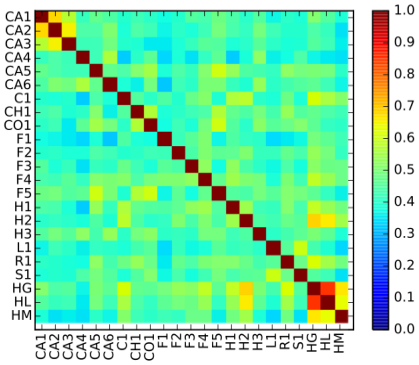


FIG. S6: Sequence similarities between frameworks – Similarity between two frameworks is measured as the fraction of common amino acids in an alignment of their two sequences. Only the library-specific part of the frameworks (Figure 1) defined in Figure S21 is considered here. In most cases, the sequence similarity is in the range 30-60%.

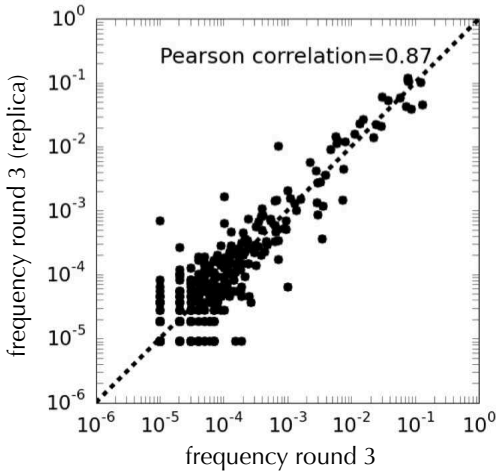


FIG. S7: Reproducibility – To assay the reproducibility of the experiments, two independent selections of the mixture of 24 libraries were performed against the DNA target and the frequencies of the sequences were compared at the third round: the high correlation between the two results indicates high reproducibility.

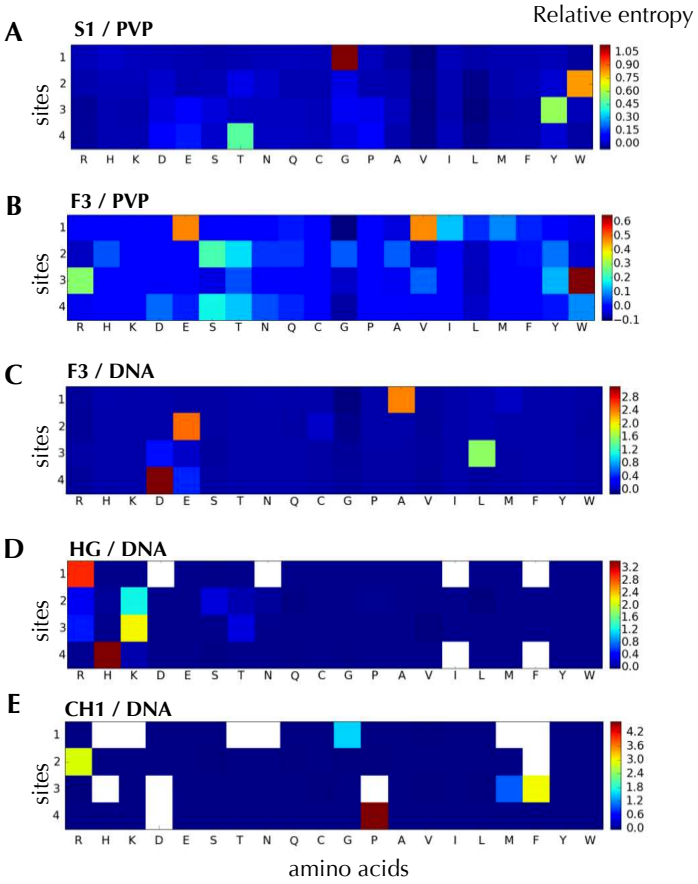


FIG. S8: Library and target specificities – Relative entropies of the different amino acids at the third round of different experiments; the relative entropy is calculated per site as $f_i^a \ln(f_i^a/q_i^a)$ where f_i^a is the frequency of amino acid a at position i in the third round and q_i^a in the initial library (round 0). **A** and **B** show that the consensus sequence is framework dependent. **B** and **C** show that it is target specific. Finally, **C**, **D** and **E** provide further evidence of framework dependency. (White squares indicate amino acid not represented in the population).

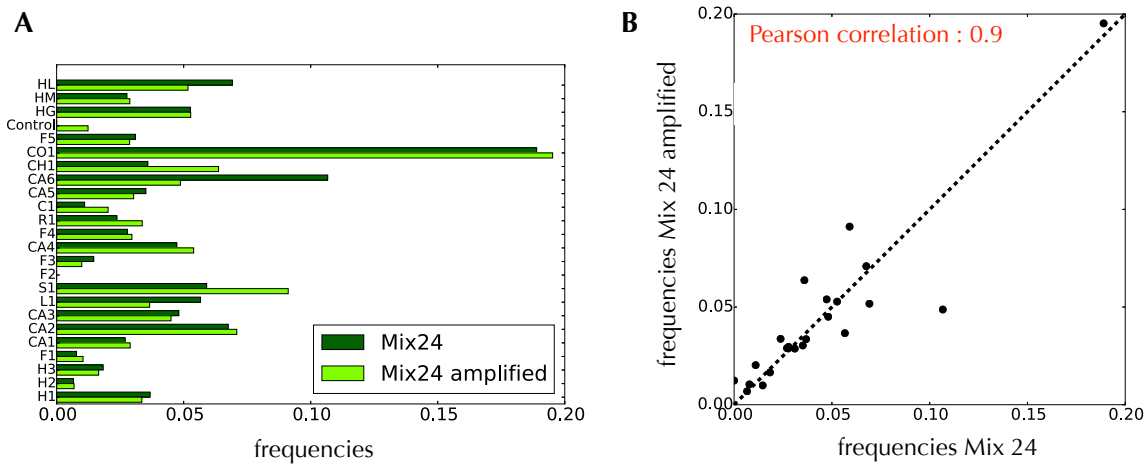


FIG. S9: Biases in amplifications – Comparison of the composition of the mixture of 24 libraries before and after amplification in absence of selection. **A.** Differences of frequencies, showing that only the S1 and CH1 libraries are enriched. **B.** Correlations between the frequencies (same data).

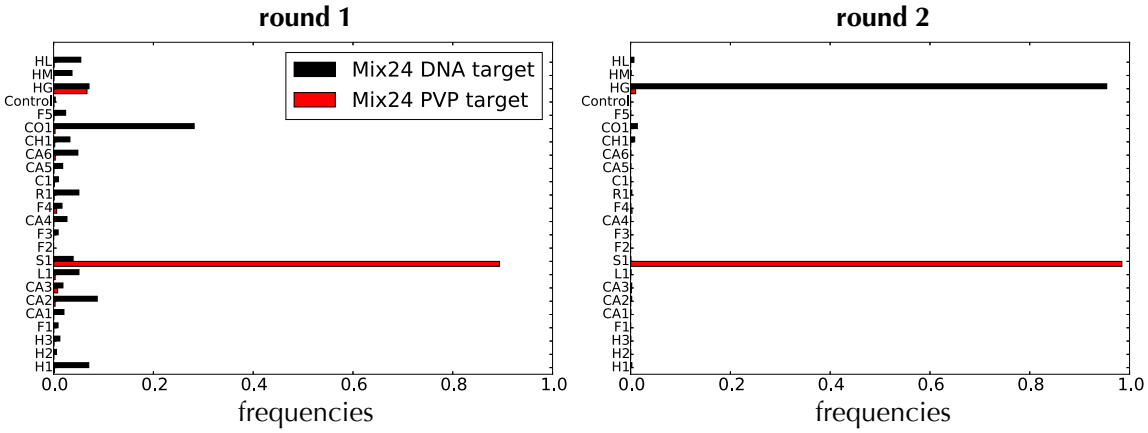


FIG. S10: Target-dependent hierarchy – Figure 2 shows that a mixture of 24 libraries selected against the DNA target is dominated by the HG framework while a mixture of 21 libraries that excludes the HG library is dominated by the CH1 framework. As shown in this figure, when the same mixture of 24 libraries is selected against the PVP target, a different framework, the S1 framework, dominates (consistently, it also dominates when screening the mixture of 21 libraries, which includes S1).

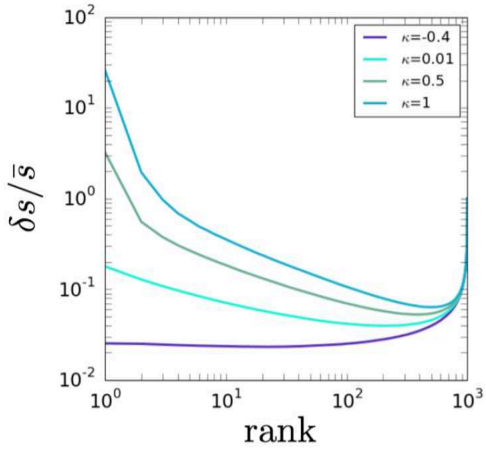


FIG. S11: Variations in the values of extreme selectivities – When sampling N random variables from the extreme probability density $f_\kappa(x)$ given by Eq. (4), the value s_r of the variable of rank r is distributed with a mean $\bar{s}_r$ and standard deviation δs_r. The ratio $\delta s_r/\bar{s}_r$ is largest for the very top sequences, as shown here based on numerical simulations. This observation is consistent with deviations of the data from a power law observed for the very top selectivities even when $\kappa > 0$ and when the overall fit with an extreme value distribution is good (Figures 4A-B).

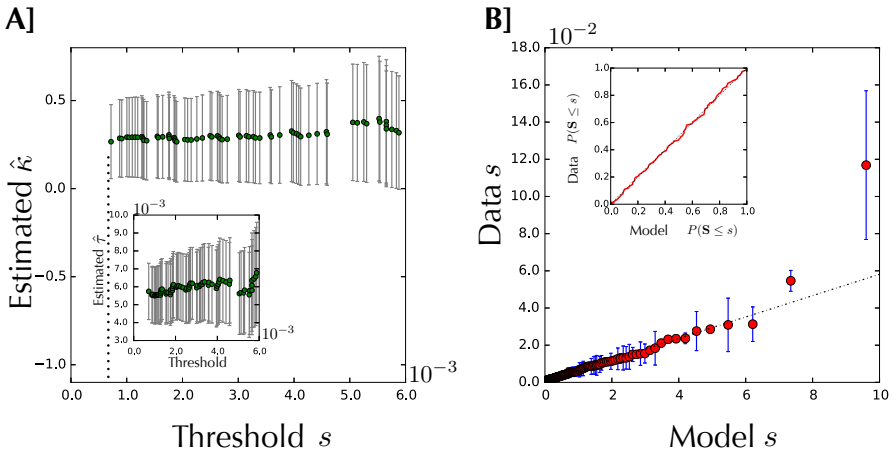


FIG. S12: EVT analysis for the selection of the HG library against the DNA target (data shown in Figure 3B). A fit of the general model gives $\kappa = 0.26 \pm 0.21$, $\tau = 5.7 \times 10^{-3} \pm 1.5 \times 10^{-3}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 8 \times 10^{-3} \pm 1.4 \times 10^{-3}$; the exponential model is excluded with a p-value 1.4×10^{-3}, in favor of $\kappa > 0$. Note that the threshold $s^* = 10^{-3}$ above which the fit is stable and good is much below the value of the selectivity above which a power law is observed in Figure 3B, of the order of $s = 10^{-2}$.

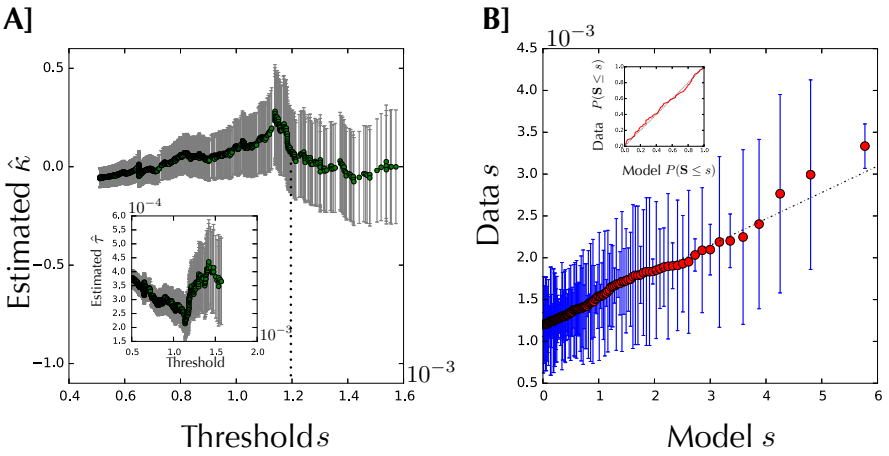


FIG. S13: EVT analysis for the selection of the F3 library against the PVP target (data shown in Figure 3D). A fit of the general model gives $\kappa = 0.07 \pm 0.21$, $\tau = 3.1 \times 10^{-4} \pm 8 \times 10^{-5}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 3.4 \times 10^{-4} \pm 6 \times 10^{-5}$; the exponential model is excluded with a p-value 0.75, which is non significant. This data is therefore consistent with an exponential model $\kappa = 0$.

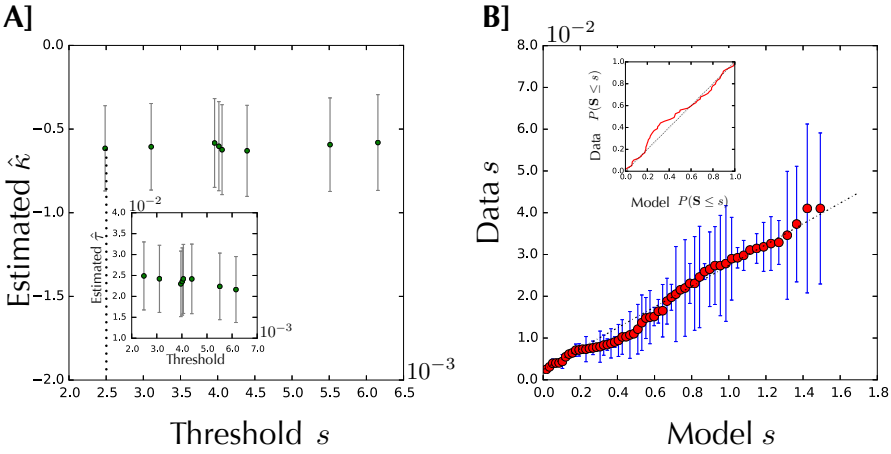


FIG. S14: EVT analysis for the selection of the CH1 library against the DNA target (data shown in Figure 3C). A fit of the general model gives $\kappa = -0.62 \pm 0.25$, $\tau = 2.5 \times 10^{-2} \pm 8 \times 10^{-3}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 1.5 \times 10^{-2} \pm 4 \times 10^{-3}$; the exponential model is excluded with a p-value 10^{-2}, in favor of $\kappa < 0$.

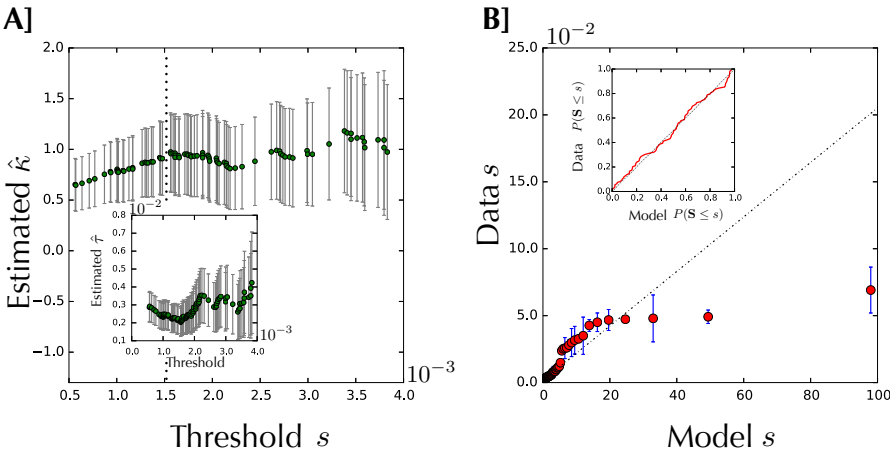


FIG. S15: EVT analysis for the selection of the F3 library against the DNA target (data not shown in the main text). A fit of the general model gives $\kappa = 0.97 \pm 0.38$, $\tau = 2 \times 10^{-3} \pm 8 \times 10^{-4}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 7.5 \times 10^{-3} \pm 10^{-3}$; the exponential model is excluded with a p-value $< 10^{-4}$, in favor of $\kappa > 0$.

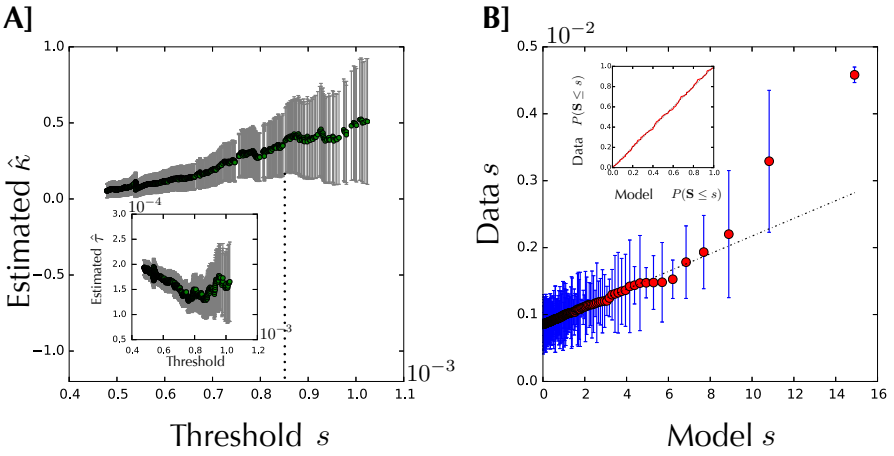


FIG. S16: EVT analysis for the selection of the N1 library in the mixture of 24 libraries against the PVP target (data not shown in the main text). A fit of the general model gives $\kappa = 0.38 \pm 0.21$, $\tau = 1.3 \times 10^{-4} \pm 3 \times 10^{-5}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 2.2 \times 10^{-4} \pm 3 \times 10^{-5}$; the exponential model is excluded with a p-value $< 10^{-4}$, in favor of $\kappa > 0$.

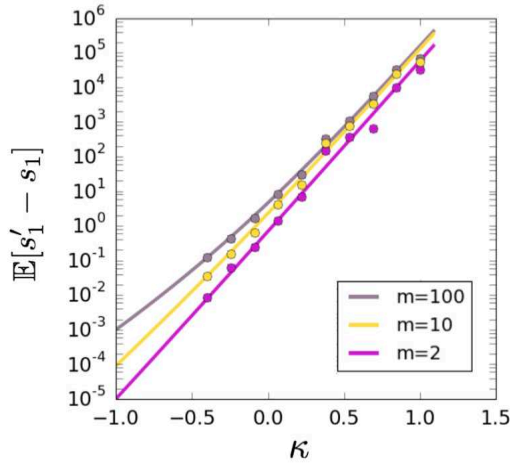


FIG. S17: Scaling of the best binder with the library size – To estimate the gain that sampling m times more the same library may provide as a function of the shape parameter κ, we show here the expected difference $\mathbb{E}[s_1' - s_1]$ between the maximum s_1' of mN samples drawn with probability density $f_\kappa(x)$ from Eq. (4) and the maximum s_1 of N sub-samples. The plain lines are based on Eq. (S8) with $\tau = 1$ and $N = 10^5$ and the dots are the results of numerical simulations (averaged over many draws), showing a good agreement between the two. Note how $\mathbb{E}[s_1' - s_1]$ depends more strongly on κ than on m.

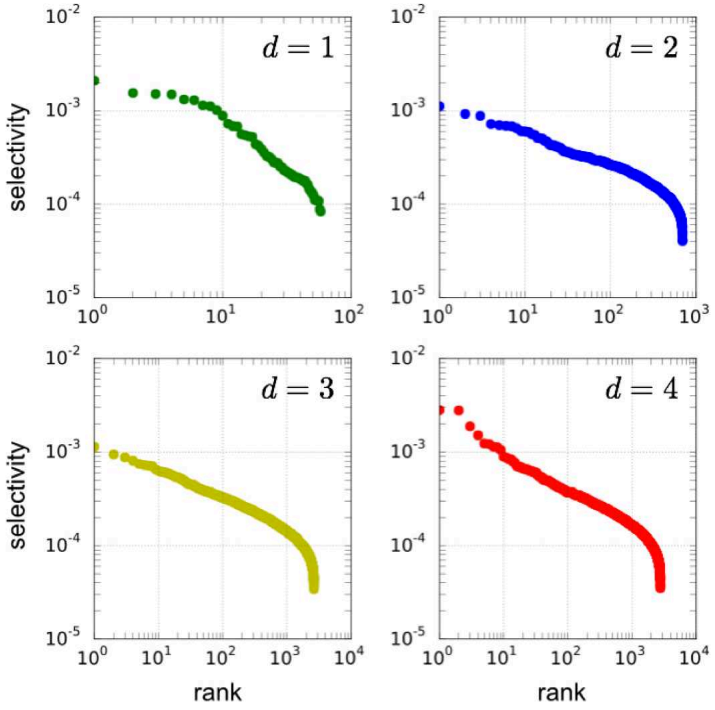


FIG. S18: Stability of the shape parameter κ for non-random sub-samples of a same library – To test whether non-random sub-libraries may be expected to be described by the same shape parameter as the library from which they originate, we consider here the results of the selection of library S1 against PVP (Figure 3A), for which the consensus CDR3 has amino acids sequence GWYT and we make four non-overlapping sub-libraries consisting of sequences with CDR3 at distance $d = 1$ to 4 from this consensus, where the distance just counts the number of amino acid differences (number of mutations). This figure shows the selectivity versus the rank of the sequences in these sub-libraries. An EVT analysis indicates that $\kappa(d = 1) = 0.33 \pm 0.39$, $\kappa(d = 2) = 0.40 \pm 0.26$, $\kappa(d = 3) = 0.30 \pm 0.23$, $\kappa(d = 4) = 0.53 \pm 0.22$: all these values are comparable to the value $\kappa = 0.44 \pm 0.22$ of the shape parameter for the full library.

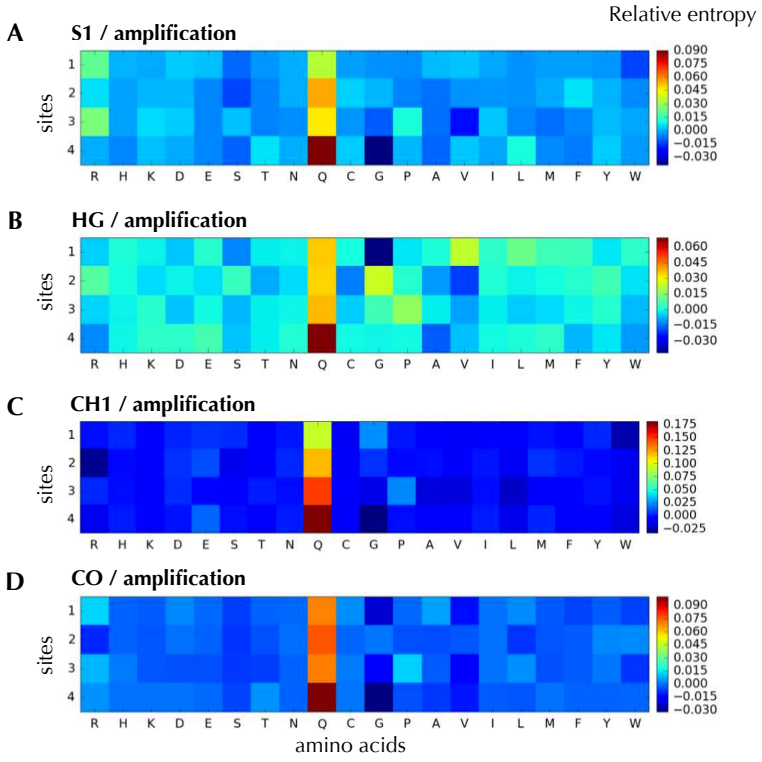


FIG. S19: Amplification bias – Relative entropy between frequencies of CDR3 sequences before and after amplification without selection, showing an enrichment in glutamine (represented by the letter Q). The results presented in the paper exclude sequences with an amber codon, which is responsible for this effect (see supplementary experimental methods), but, in most experiments with selection, glutamine does not appear in the selected consensus sequence and considering the amber code as coding for an amino acid or for a stop codon has no incidence on the conclusions.

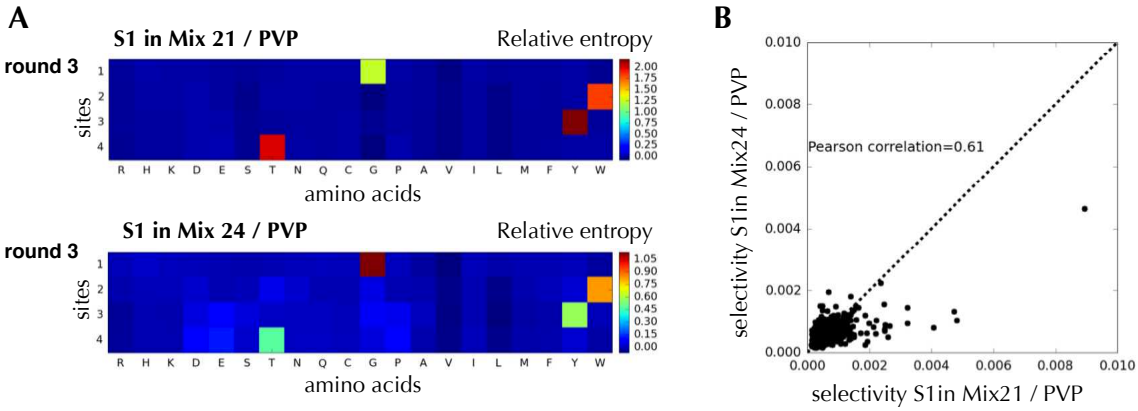


FIG. S20: Reproducibility of selections against the PVP target – The results of two experiments of selection against the PVP target, one starting from a mixture of 24 libraries and the other from a subset of 21 libraries, which each are dominated by the S1 library, not only lead to an identical consensus sequence (panel A) but to reproducible results by EVT analysis (Table SIII). In this case, not only are the initial populations different, but also potentially the targets since the experiments were performed 1.5 year apart and PVP is subject to aging: this may explain the imperfect correlations between frequencies (panel B; by contrast, the selection of the F3 library against PVP was performed at the same time than the selection of the mixture of 24 libraries and differences in consensus sequences cannot be due to differences of the targets in this case).

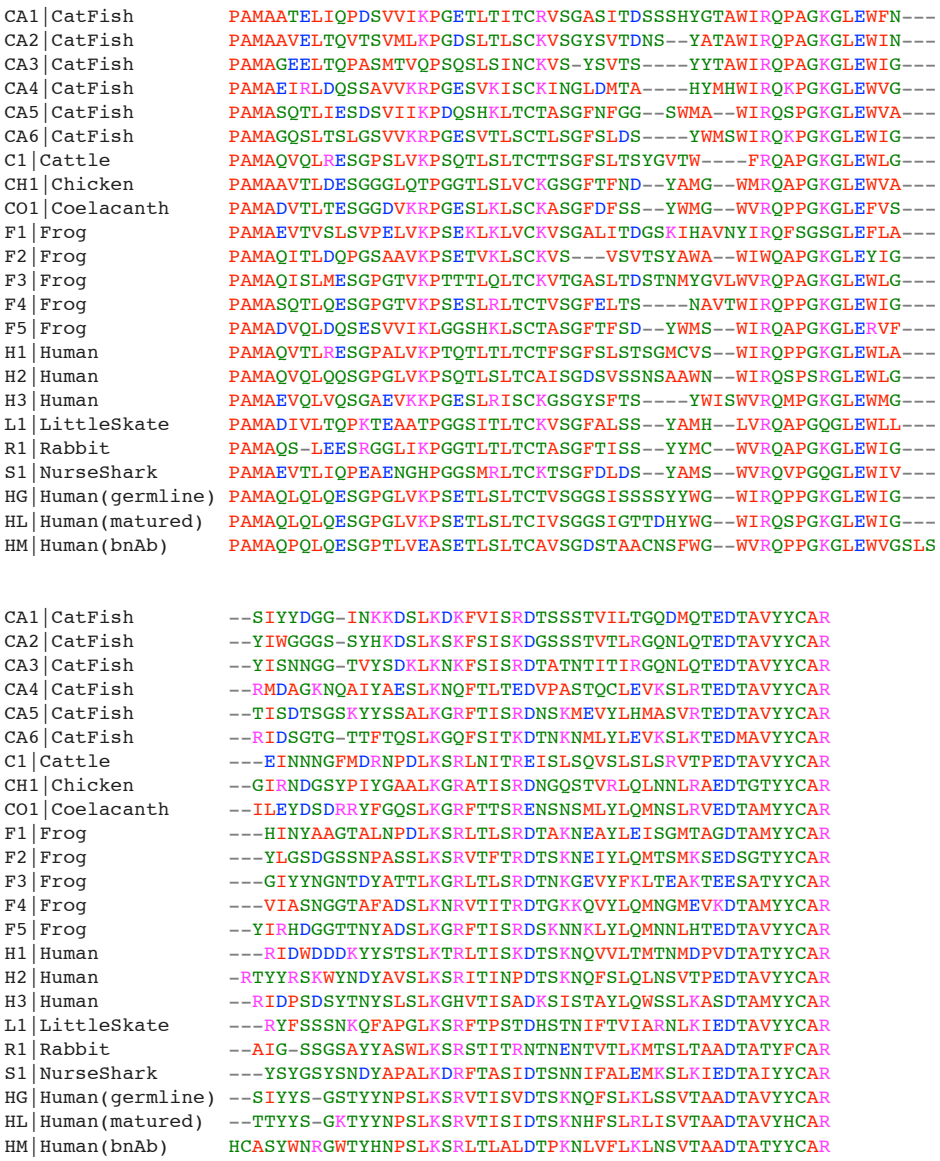


FIG. S21: Amino acid sequences of the frameworks – Multiple sequence alignment of the library-specific part of the frameworks (Figure 1). The organism from which the sequence originate is indicated.

BOYER SÉBASTIEN

ANNEXES

Table des matières

Séquences Q>0

Il n'y a quasiment pas de différences entre ces figures et celles présentées dans la partie *Resultats*, à part l'exacte valeur de l'entropie relative ou de l'information mutuelle : les erreurs de lectures présentes dans nos données (un facteur 2 à 5 entre Q>0 et Q>30) n'affectent pas ce type d'analyse.

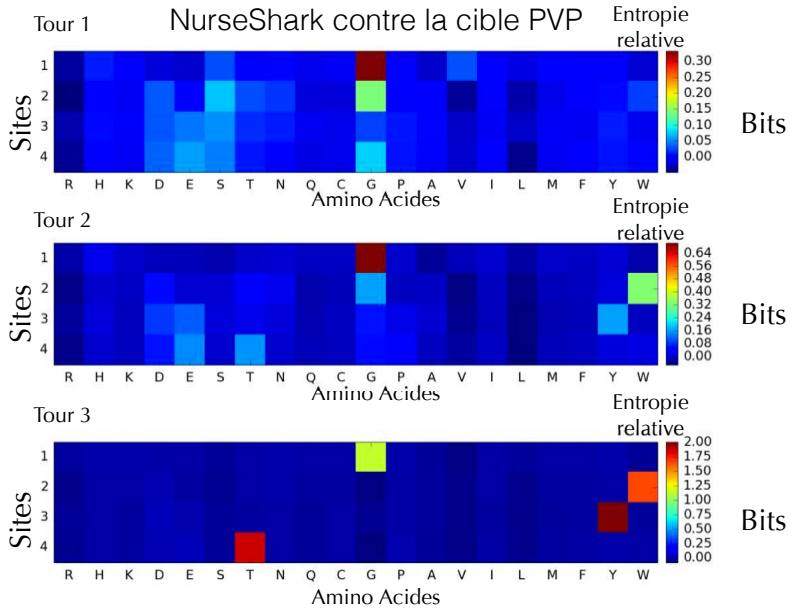


FIGURE 1: **Entropie relative de chaque sites du CDR3 au cours de la sélection.**

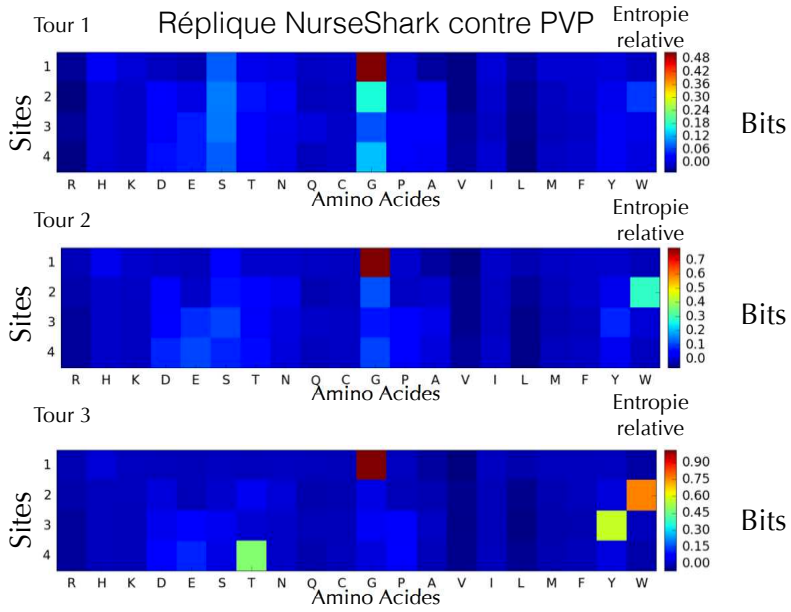


FIGURE 2: **Entropie relative de chaque sites du CDR3 au cours de la sélection.**

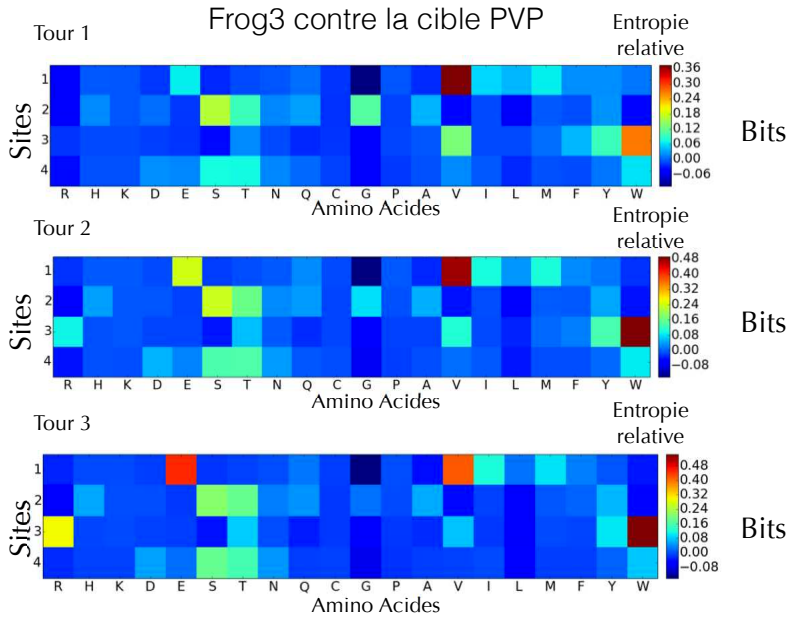


FIGURE 3: **Entropie relative de chaque sites du CDR3 au cours de la sélection.**

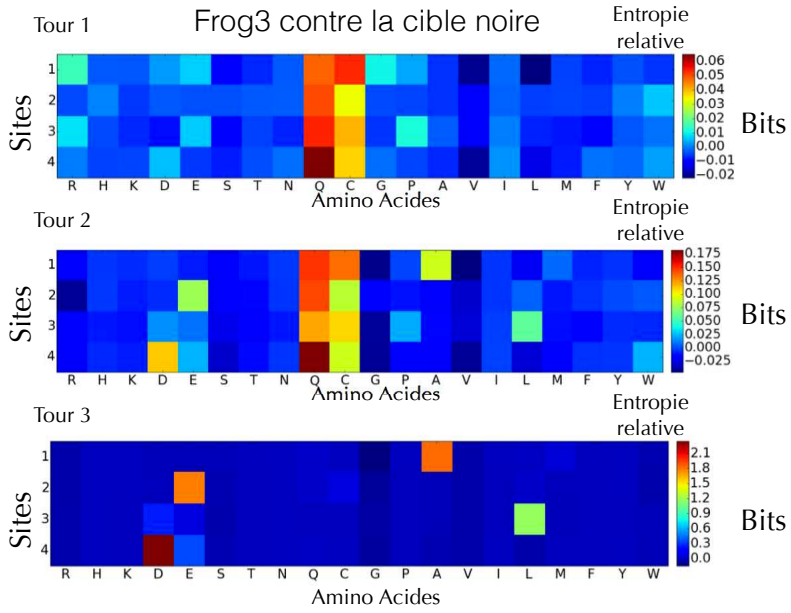


FIGURE 4: **Entropie relative de chaque sites du CDR3 au cours de la sélection.**

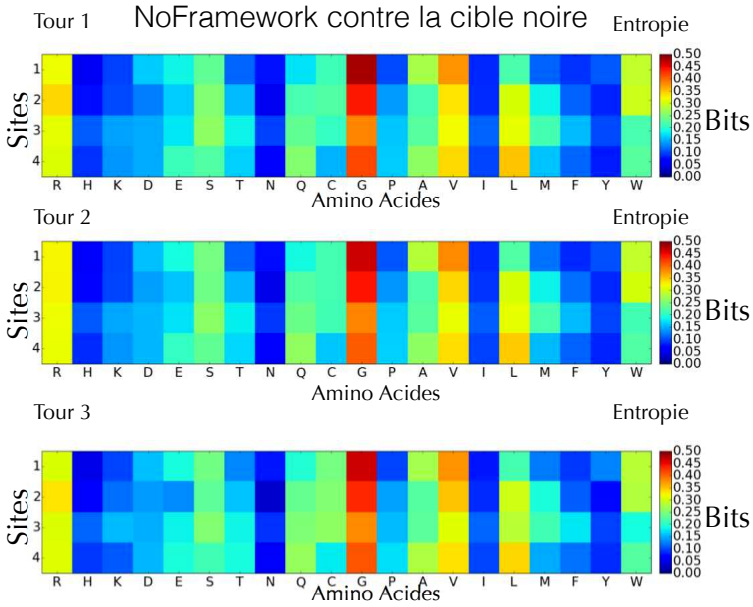


FIGURE 5: **Entropie de chaque sites du CDR3 au cours de la sélection.**

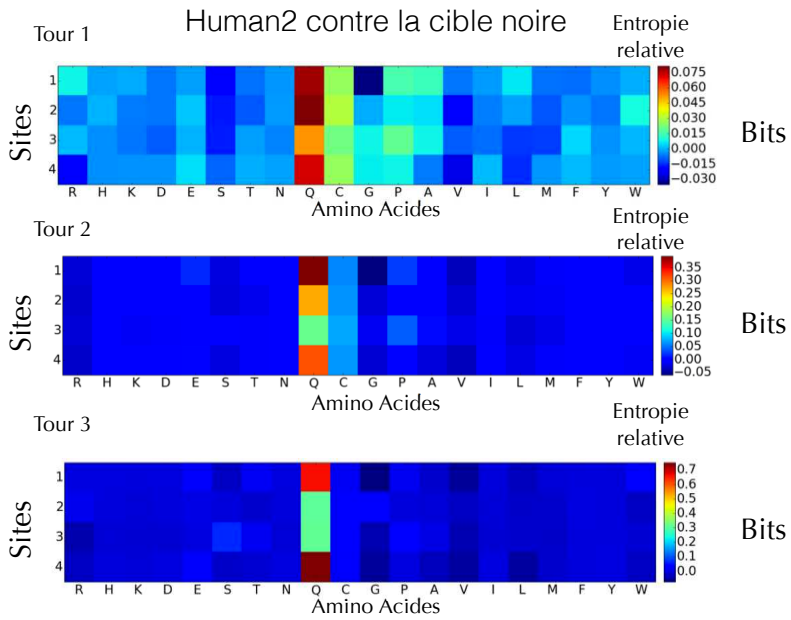


FIGURE 6: **Entropie relative de chaque sites du CDR3 au cours de la sélection.**

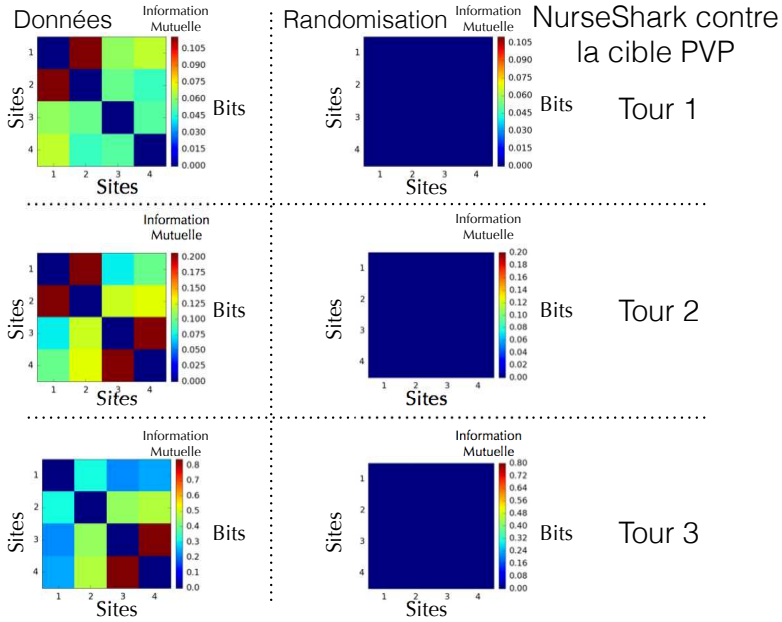


FIGURE 7: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

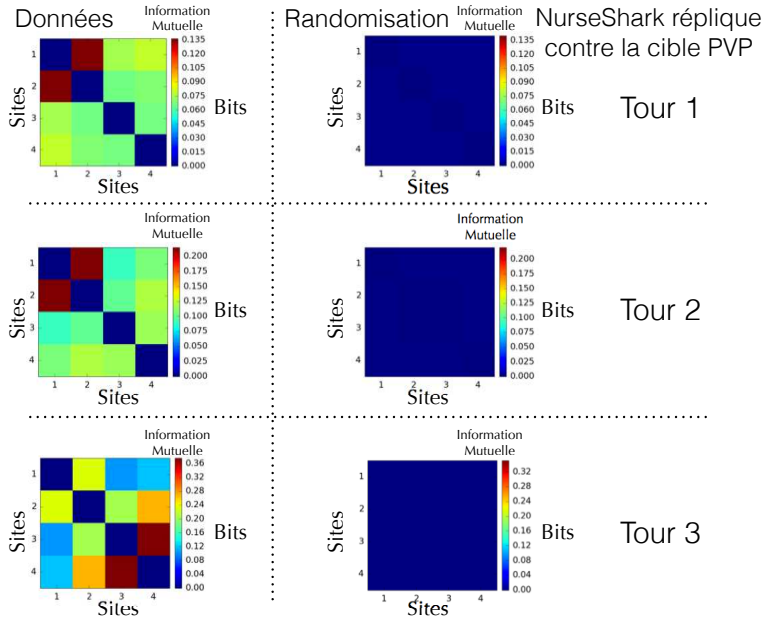


FIGURE 8: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

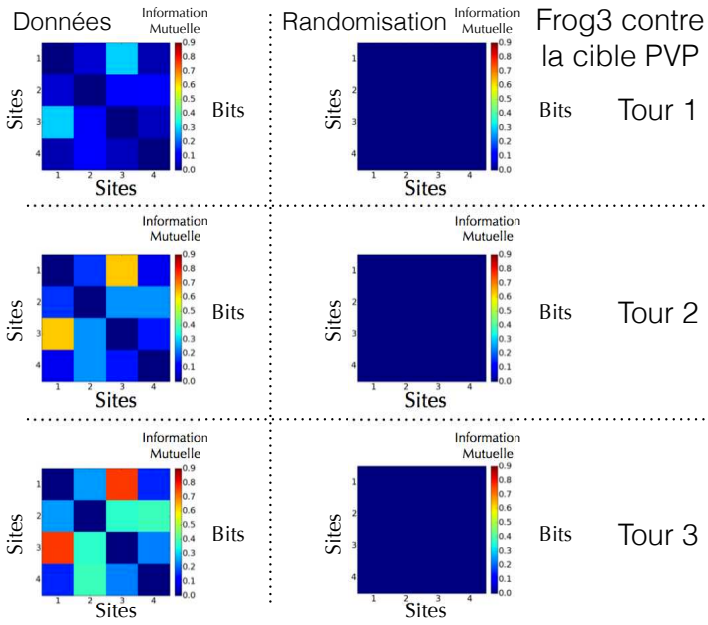


FIGURE 9: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

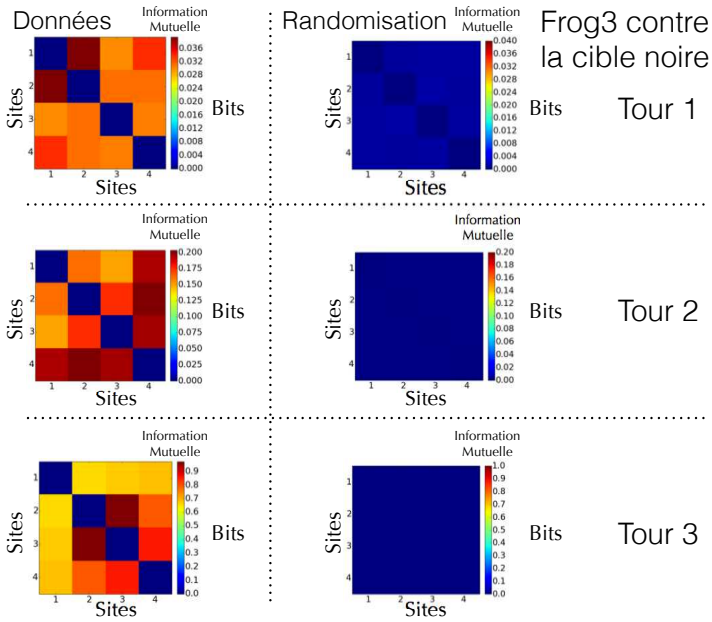


FIGURE 10: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

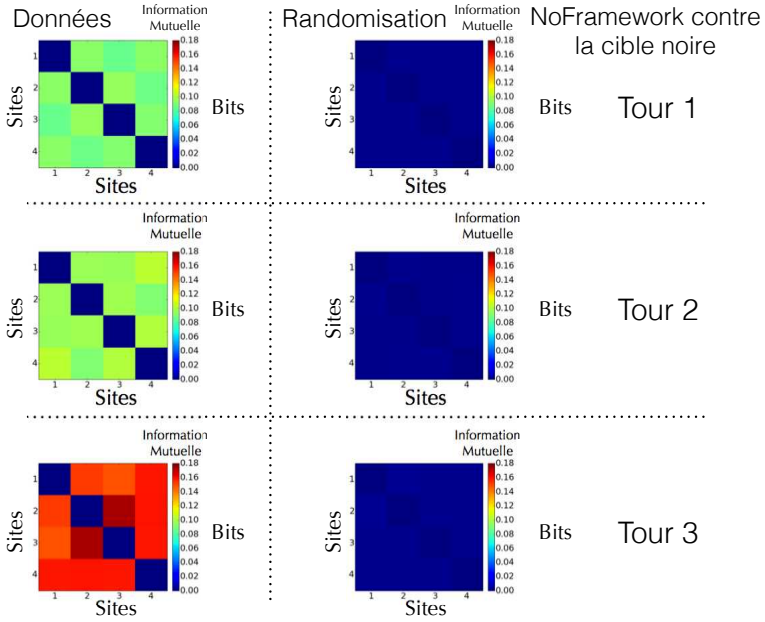


FIGURE 11: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

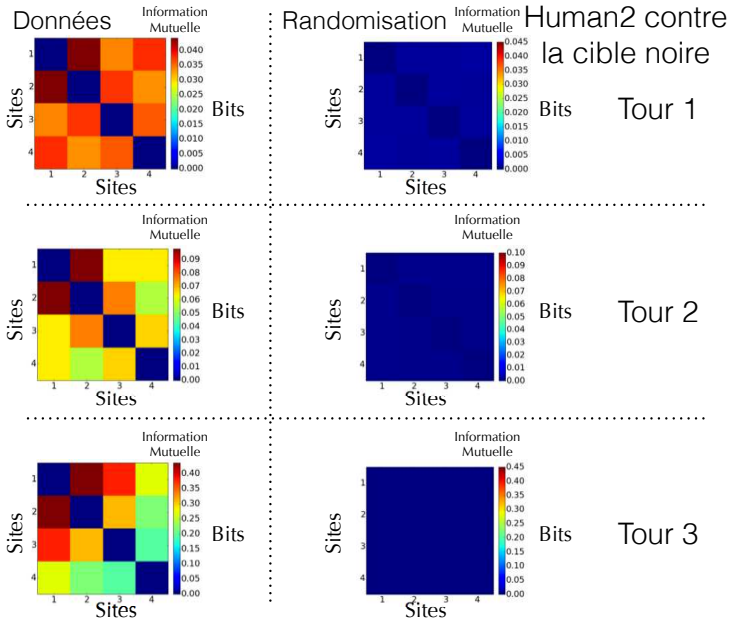


FIGURE 12: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

Séquences Mixtures Q>30

Dans certains cas de mixtures, l'analyse en terme de séquences n'a pas été présentée car la mixture de frameworks persistait avec les tours. L'analyse framework par framework est ici présentée, quand elle est possible.

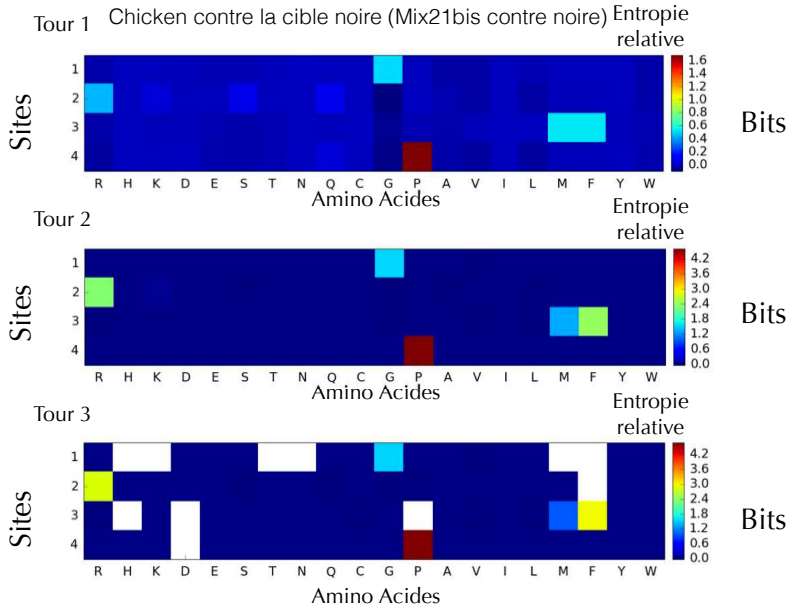


FIGURE 13: **Entropie relative de chaque sites du CDR3 au cours de la sélection.** Les cases blanches de la matrices correspondent à des entropies relatives non calculées car ces acides aminés ne sont plus présents à ces positions dans le troisième tour. Une séquence de type 'GR(MF)P' est sélectionnée.

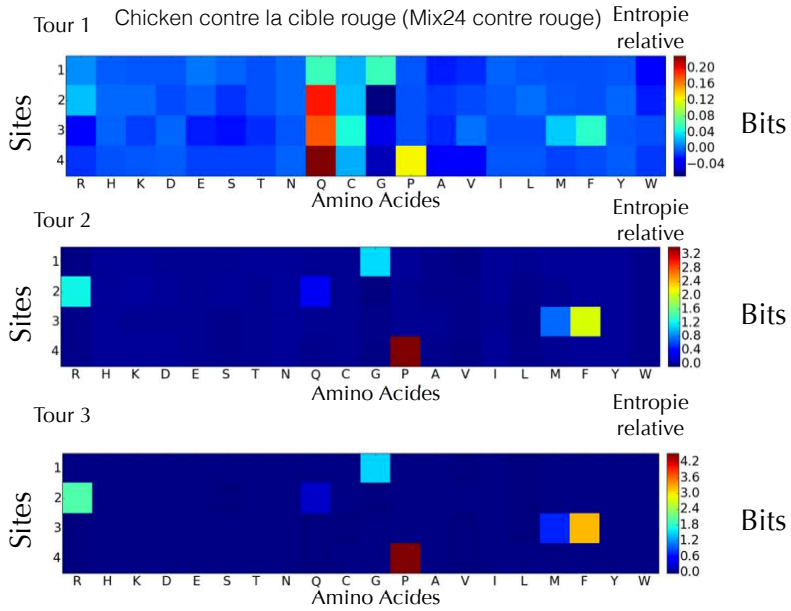


FIGURE 14: **Entropie relative de chaque sites du CDR3 au cours de la sélection.** Une séquence de type 'GR(MF)P' est sélectionnée.

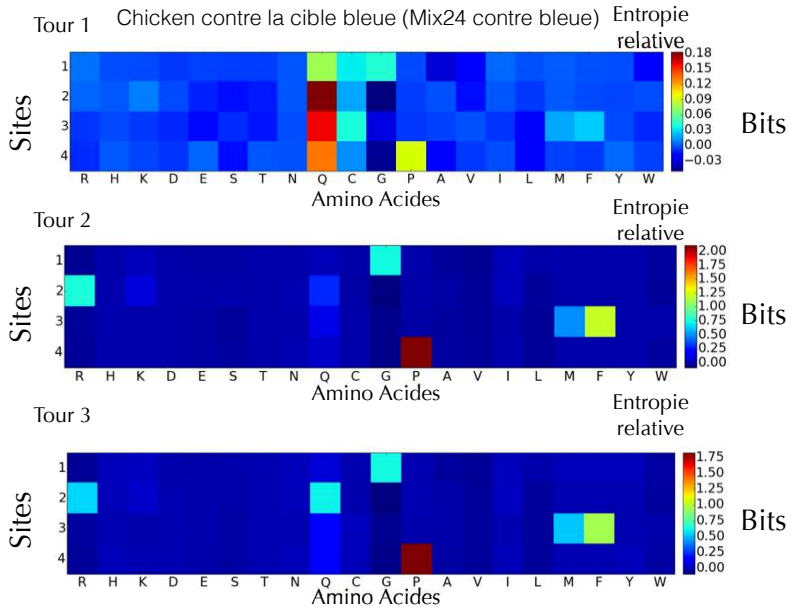


FIGURE 15: **Entropie relative de chaque sites du CDR3 au cours de la sélection.** Une séquence de type 'GR(MF)P' est sélectionnée.

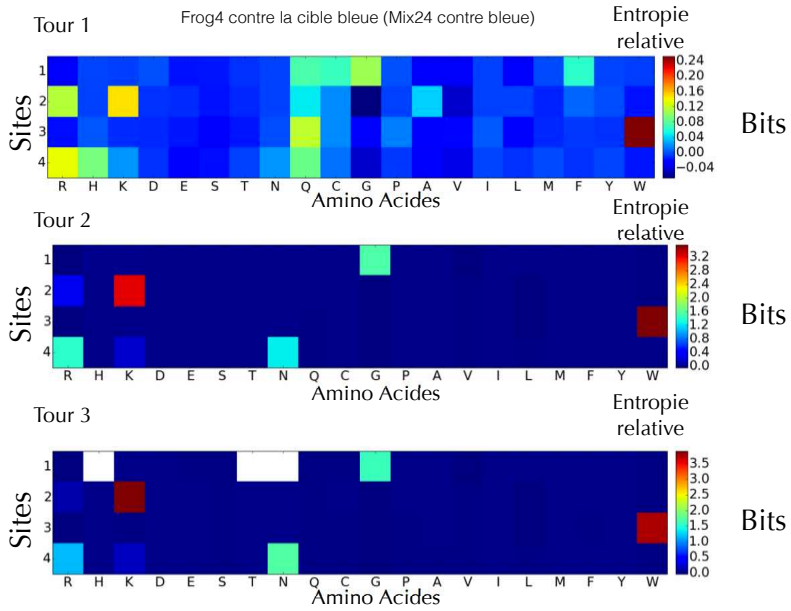


FIGURE 16: **Entropie relative de chaque sites du CDR3 au cours de la sélection.** Une séquence de type 'GKW(RN)' est sélectionnée.

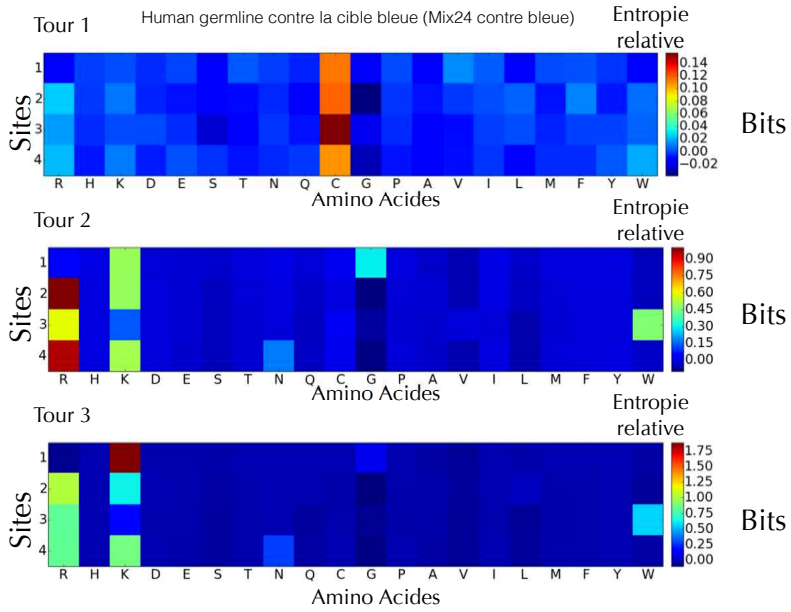


FIGURE 17: **Entropie relative de chaque sites du CDR3 au cours de la sélection.** Une séquence riche en acide aminée positifs est sélectionné : 'K(RK)(RW)(RK)'.

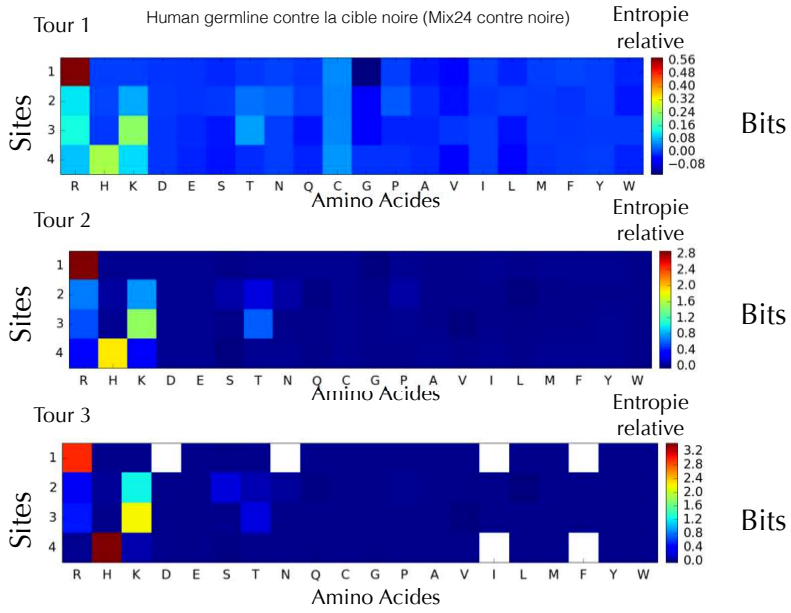


FIGURE 18: **Entropie relative de chaque sites du CDR3 au cours de la sélection.** Une séquence riche en acide aminée positifs est sélectionné : 'RKKH'. Il est remarquable de voire que même si le type d'interaction avec l'ADN est le même (c'est à dire électrostatique), la séquence consensus est quand même cible spécifique.

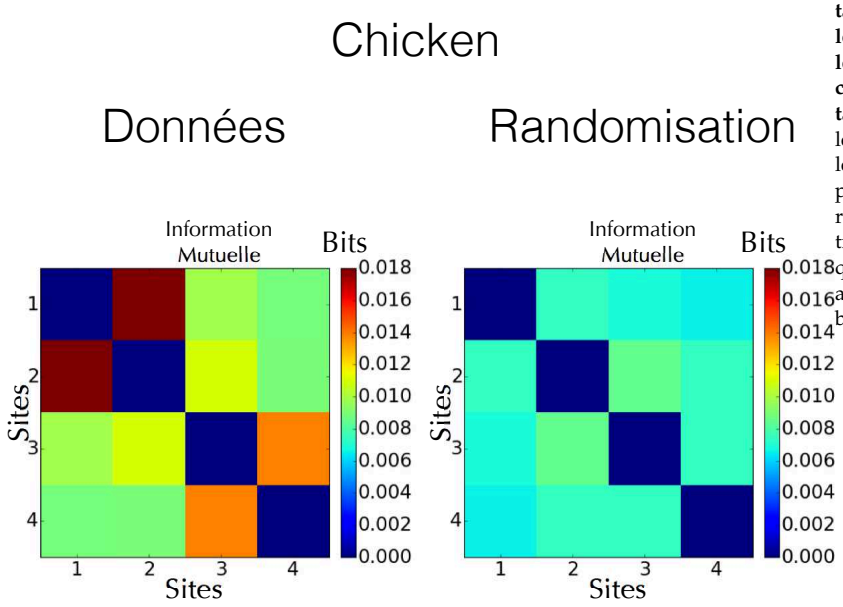


FIGURE 19: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.** Comme pour les banques initiales présentées dans le corps du manuscrit, on ne peut pas faire la différence entre corrélations réelles et corrélations dues à l'échantillonnage. On remarque tout de même que ces corrélations sont très similaires aux corrélations vues dans les autres banques initiales.

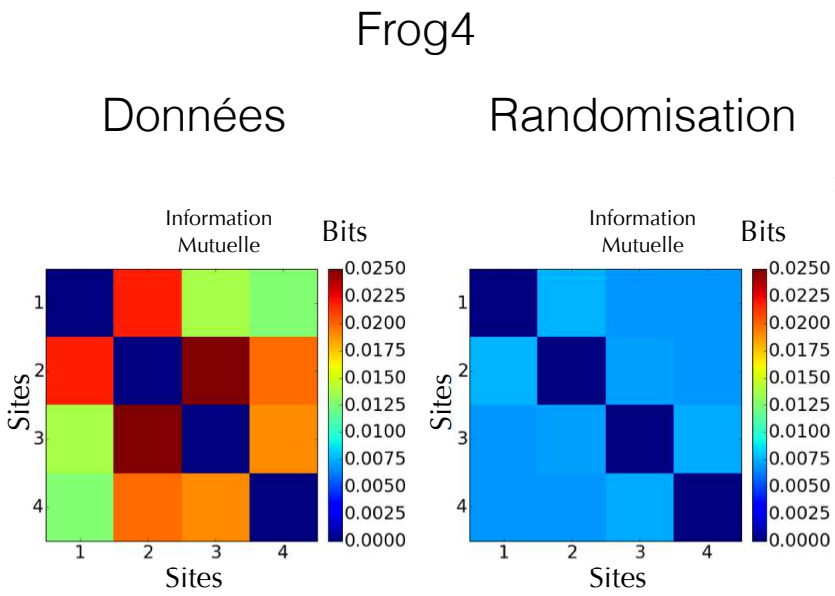


FIGURE 20: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.** Comme pour les banques initiales présentées dans le corps du manuscrit, on ne peut pas faire la différence entre corrélations réelles et corrélations dues à l'échantillonnage. On remarque tout de même que ces corrélations sont très similaires aux corrélations vues dans les autres banques initiales.

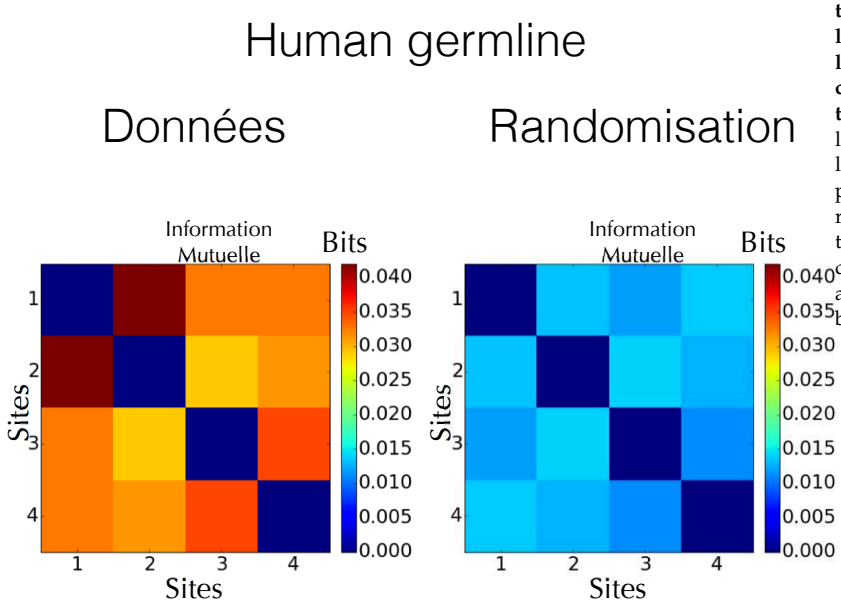


FIGURE 21: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.** Comme pour les banques initiales présentées dans le corps du manuscrit, on ne peut pas faire la différence entre corrélations réelles et corrélations dues à l'échantillonnage. On remarque tout de même que ces corrélations sont très similaires aux corrélations vues dans les autres banques initiales.

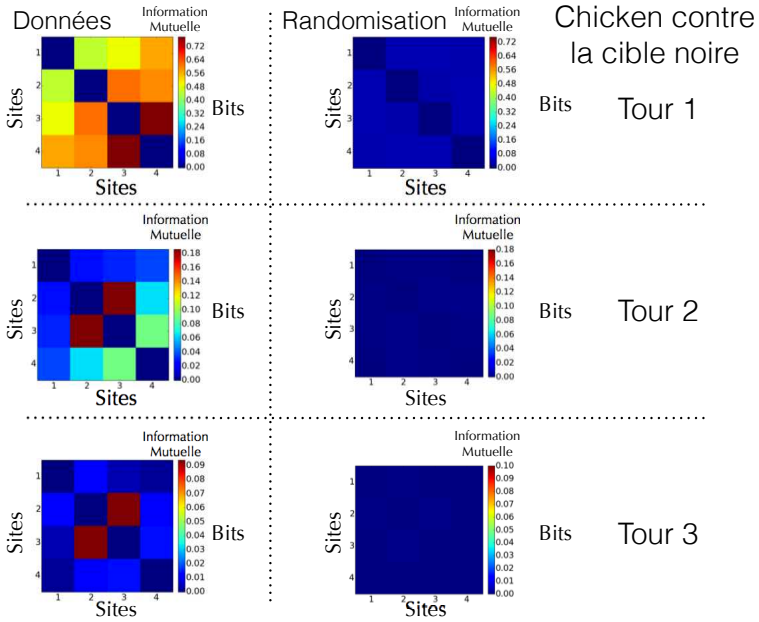


FIGURE 22: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.** Au tour 1 la banque contient encore beaucoup des corrélations de la banque initiales. Ces corrélations baissent au tour 2 et restent stables au tour 3 : les corrélations vues au tour 1 ne représentent pas la sélection. Tout les chickens contre cibles ADN présentent la même distribution de fréquences et de corrélations au tour 3, ce qui indique surement une sélection pour la partie commune des trois cibles : la tige.

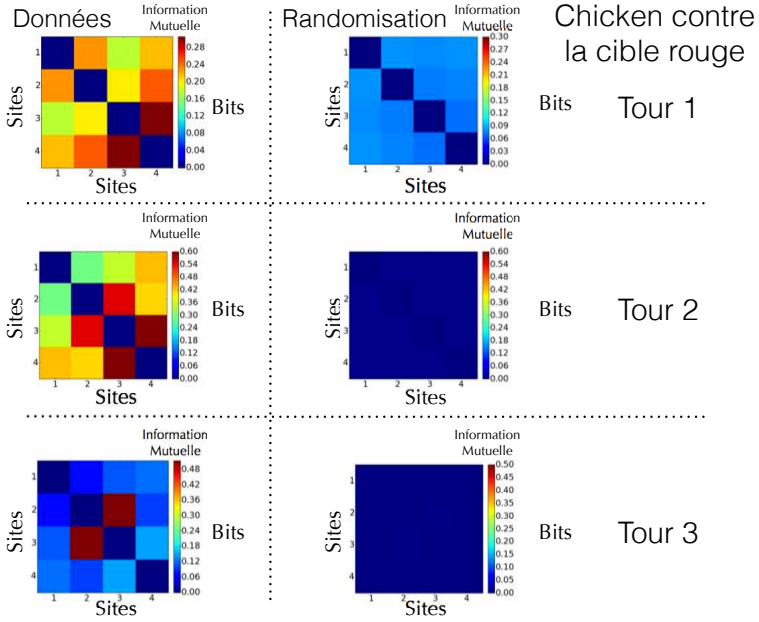


FIGURE 23: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

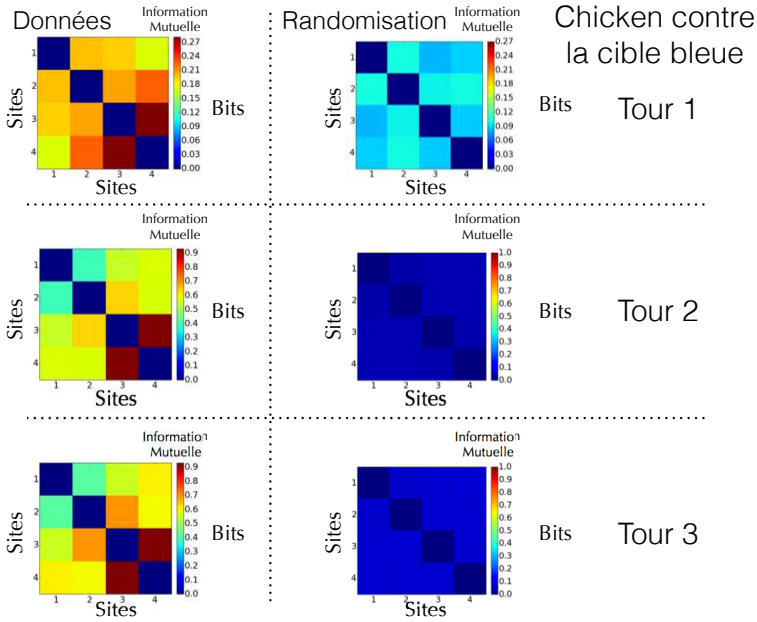


FIGURE 24: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

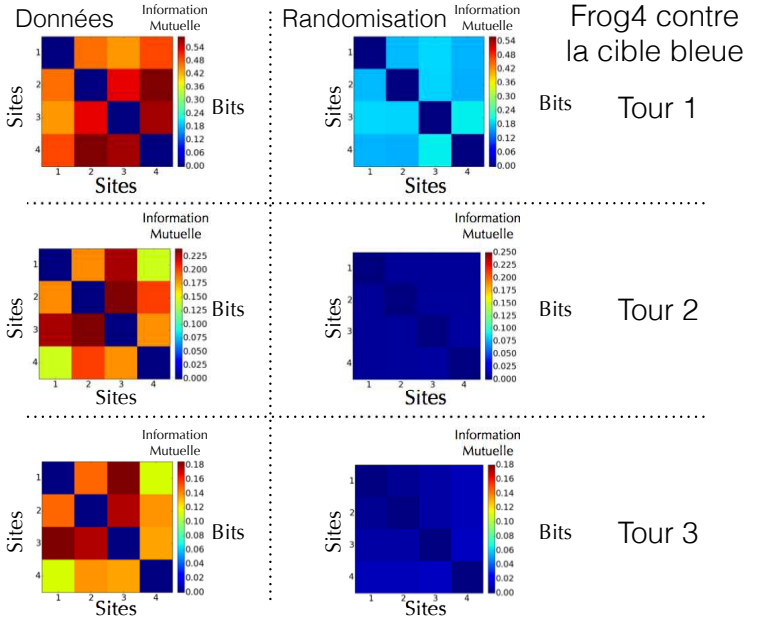


FIGURE 25: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

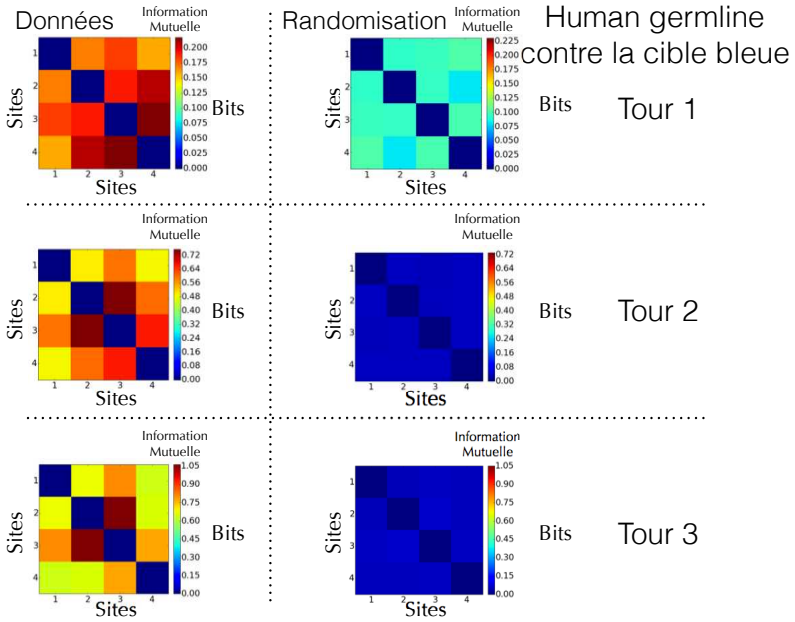


FIGURE 26: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

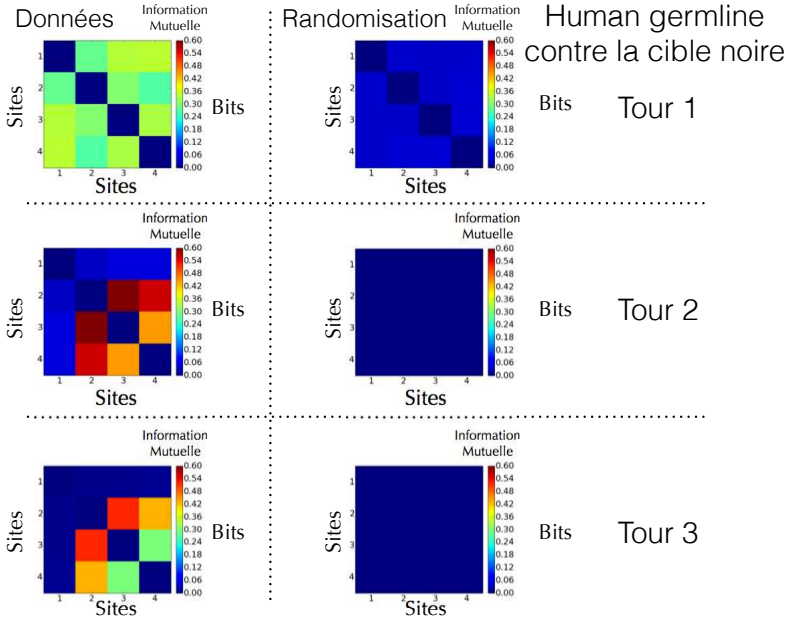


FIGURE 27: **Information mutuelle partagée entre les sites du CDR3 pour les données et un modèle respectant les fréquences des données mais sans corrélations et échantillonné au même taux que les données.**

Valeurs extrêmes Mix21 contre PVP

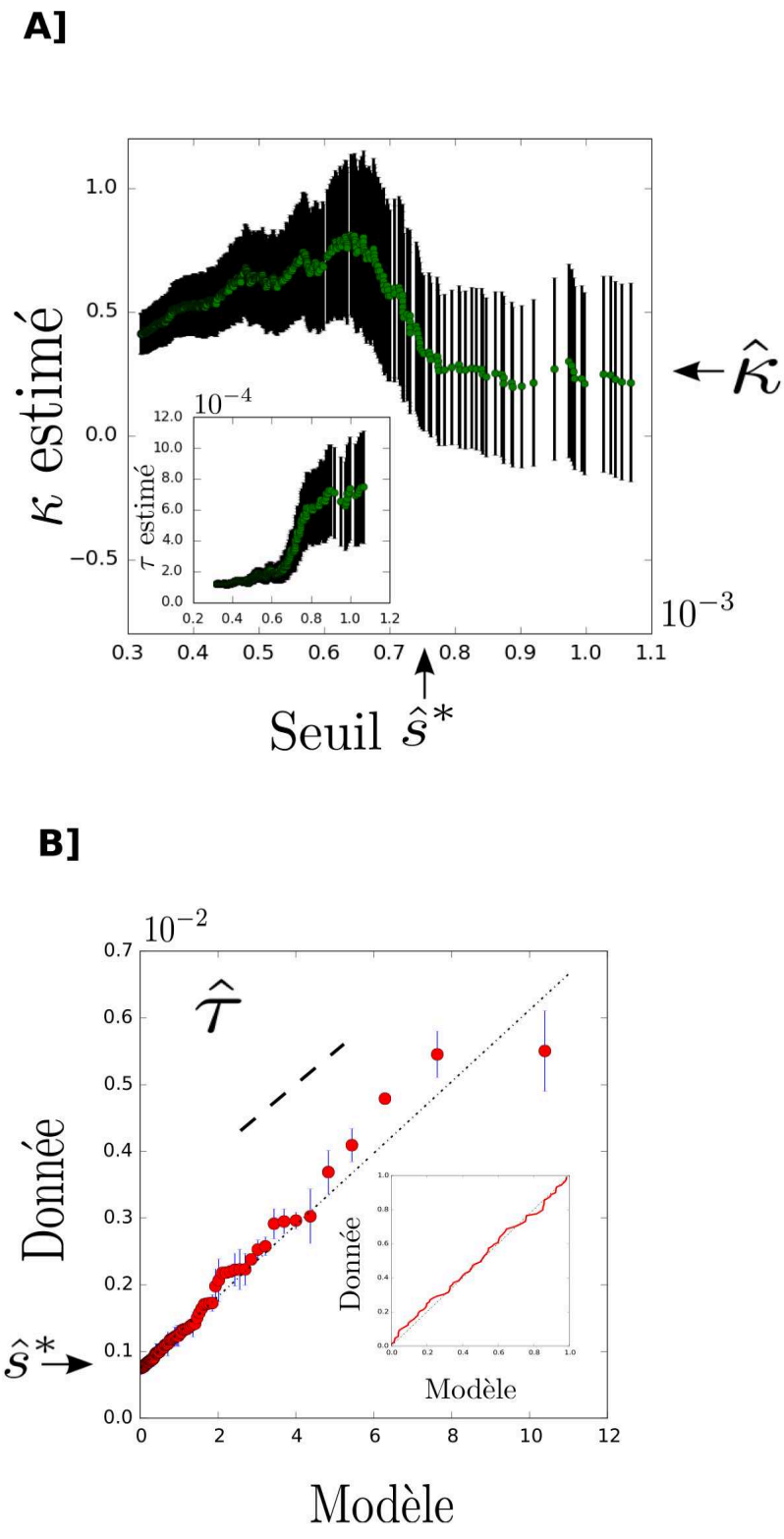


FIGURE 28: **Robustesse de l'analyse en valeurs extrêmes Mix21contre PVP Q>30** $\frac{\Delta s}{s} < 0.3$

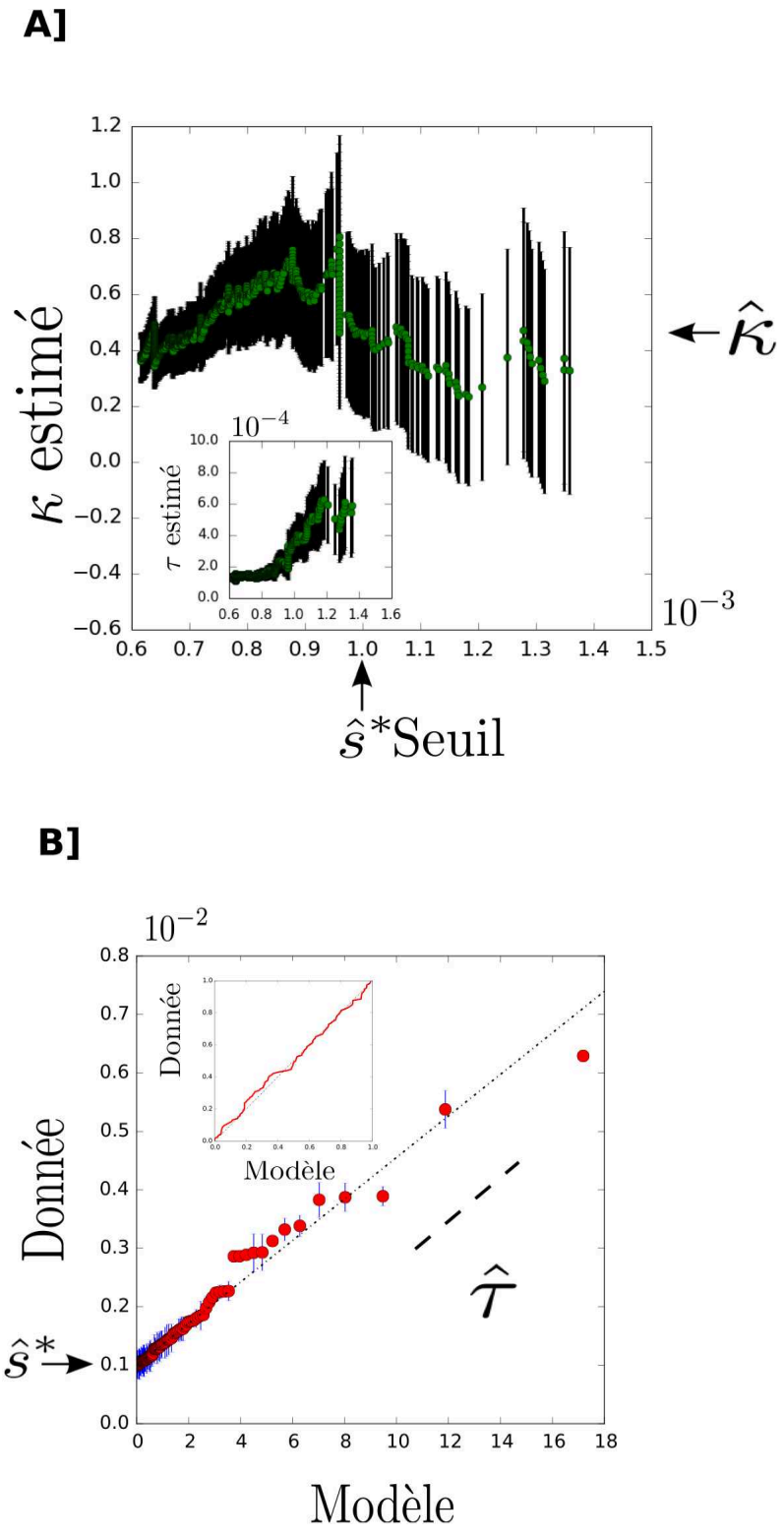


FIGURE 29: **Robustesse de l'analyse en valeurs extrêmes Mix21 contre PVP Q>30** $\Delta s < s_{noise}$

Valeurs extrêmes NoShark contre PVP

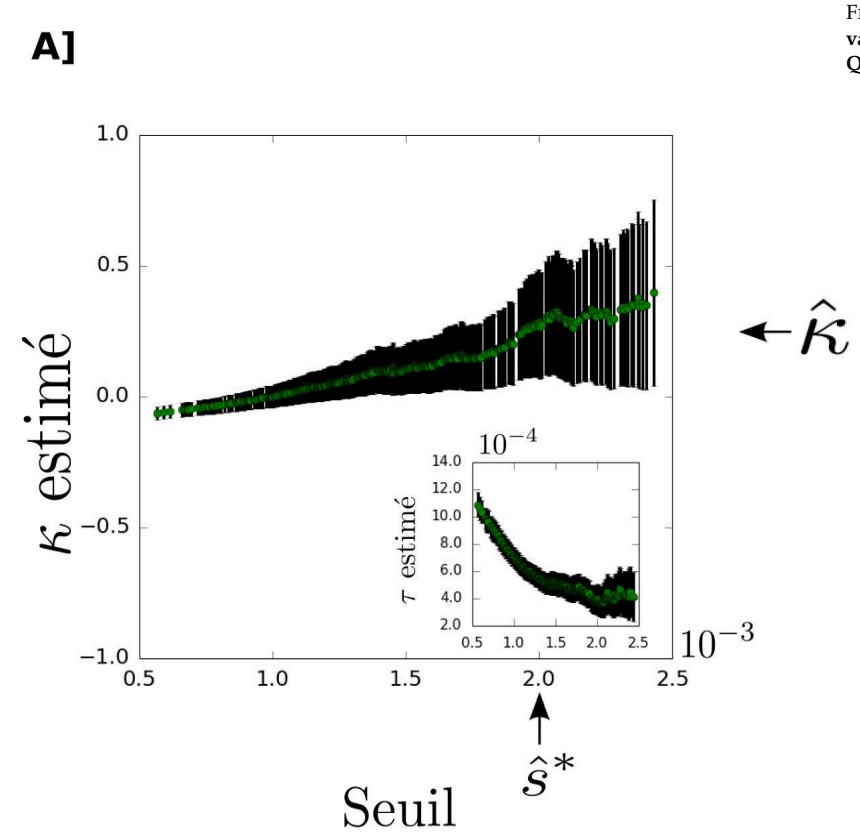
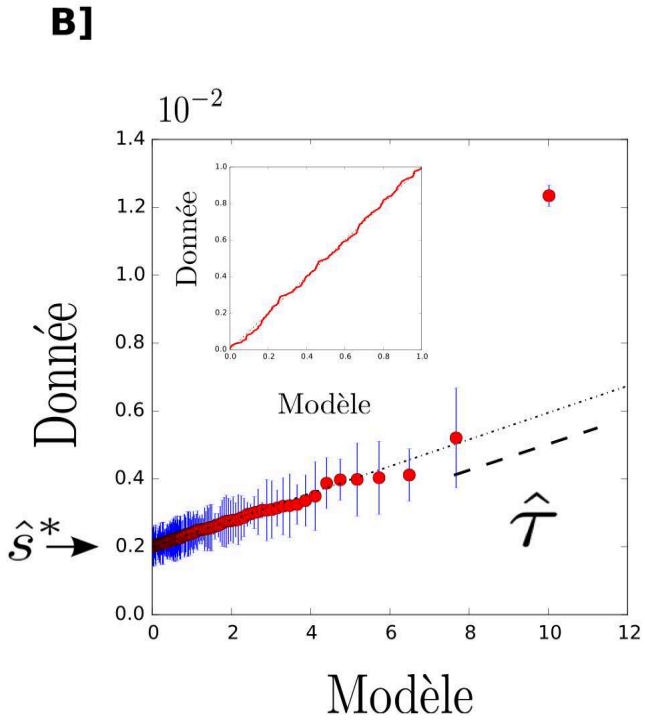


FIGURE 30: **Robustesse de l'analyse en valeurs extrêmes NoShark contre PVP Q>30** $\frac{\Delta s}{s} < 0.3$



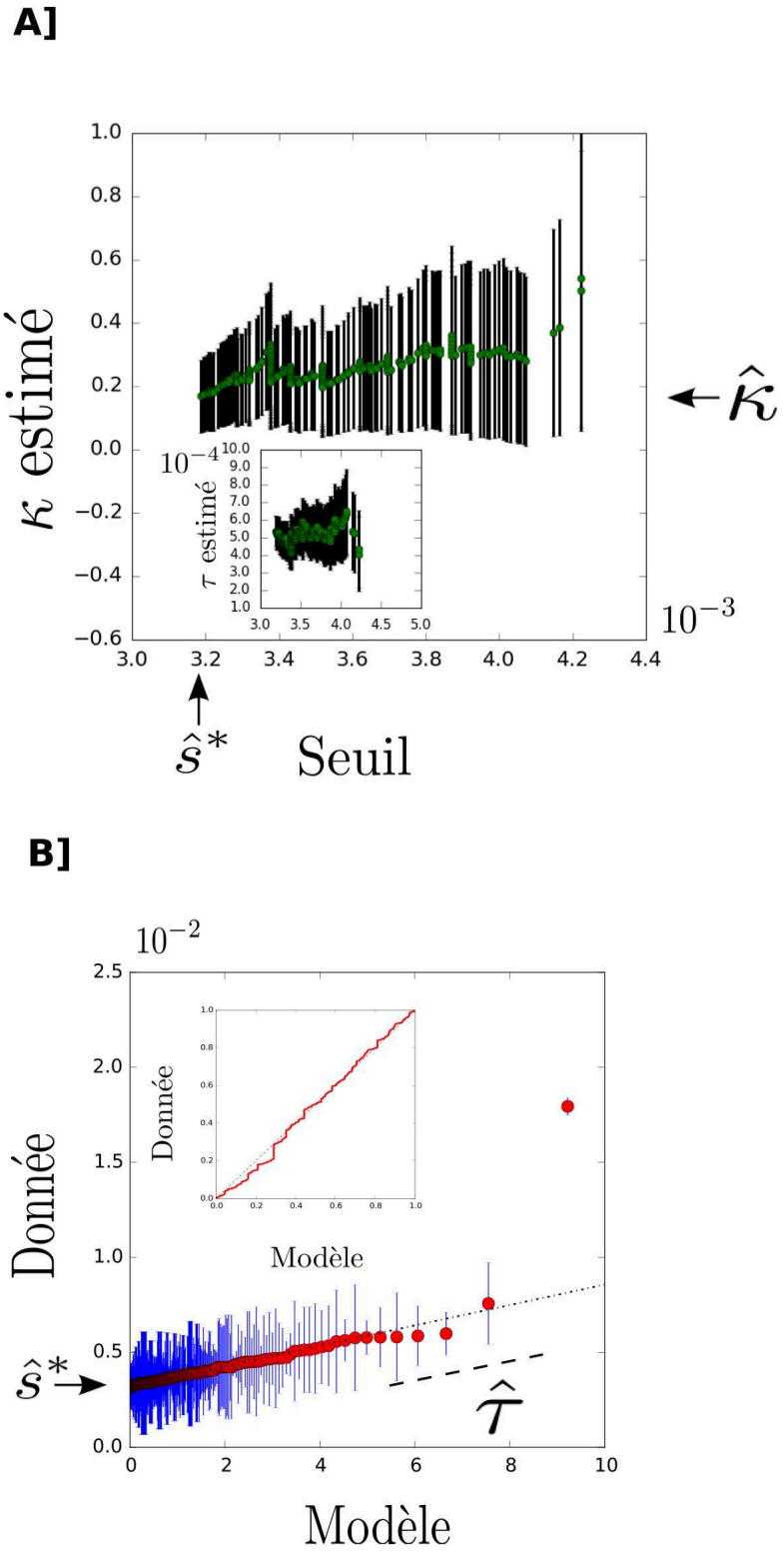


FIGURE 31: **Robustesse de l'analyse en valeurs extrêmes NoShark contre PVP Q>30** $\Delta s < s_{noise}$

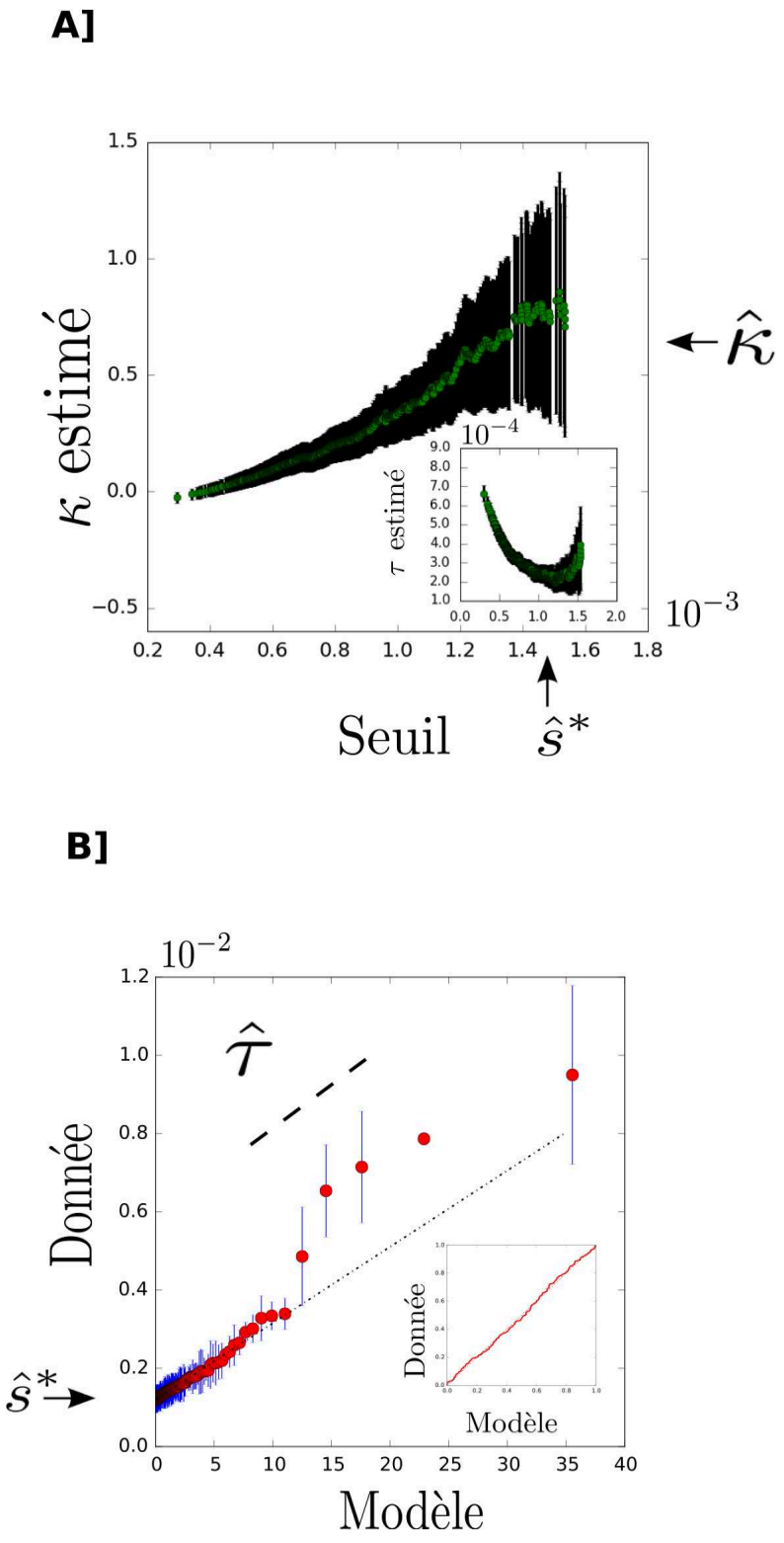


FIGURE 32: **Robustesse de l'analyse en valeurs extrêmes NoShark contre PVP Q>o** $\frac{\Delta s}{s} < 0.3$

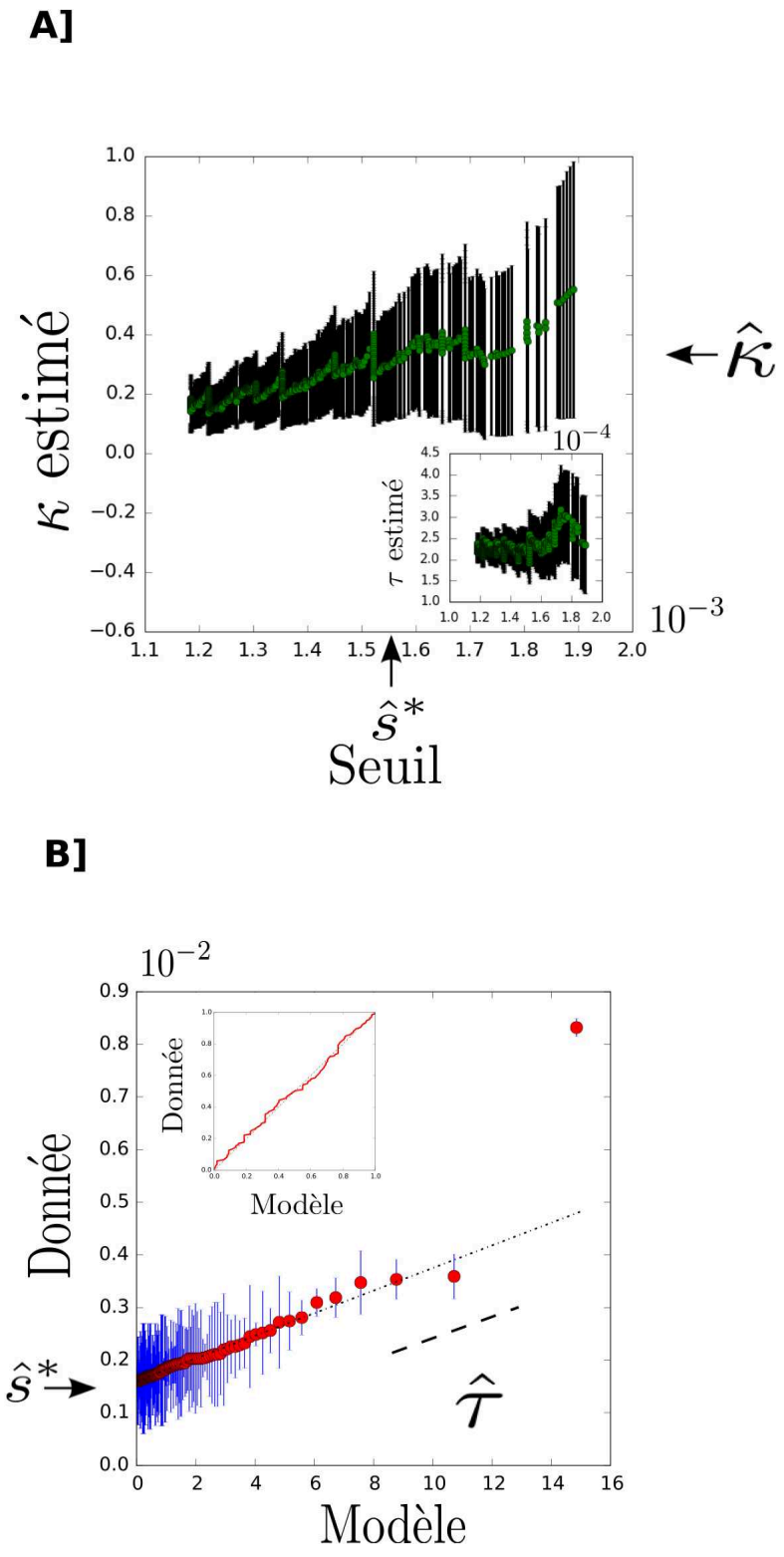


FIGURE 33: **Robustesse de l'analyse en valeurs extrêmes NoShark contre PVP Q>o** $\Delta s < s_{noise}$

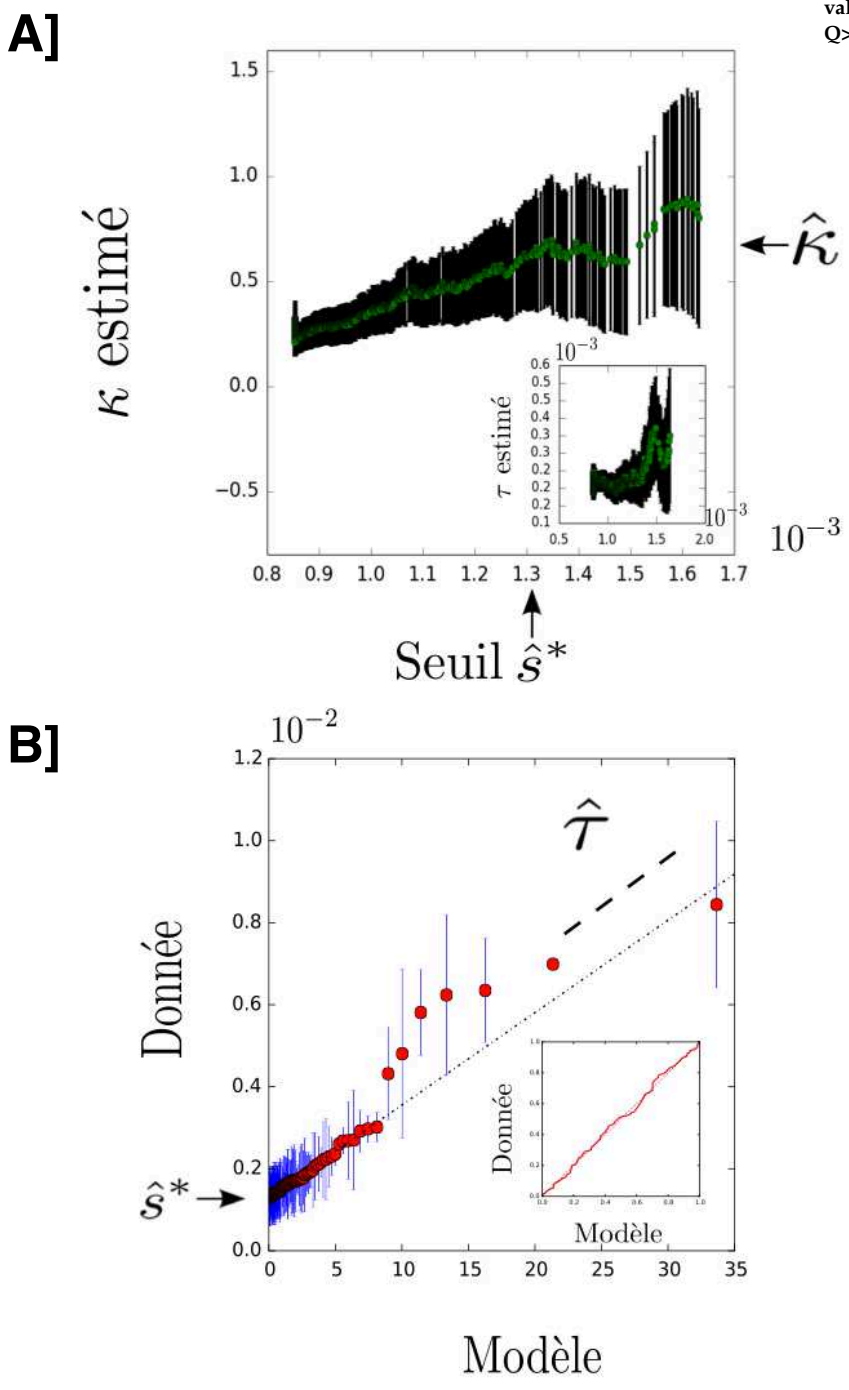


FIGURE 34: **Robustesse de l'analyse en valeurs extrêmes NoShark contre PVP Q>o** $n > 10$

Valeurs extrêmes Frog3 contre PVP

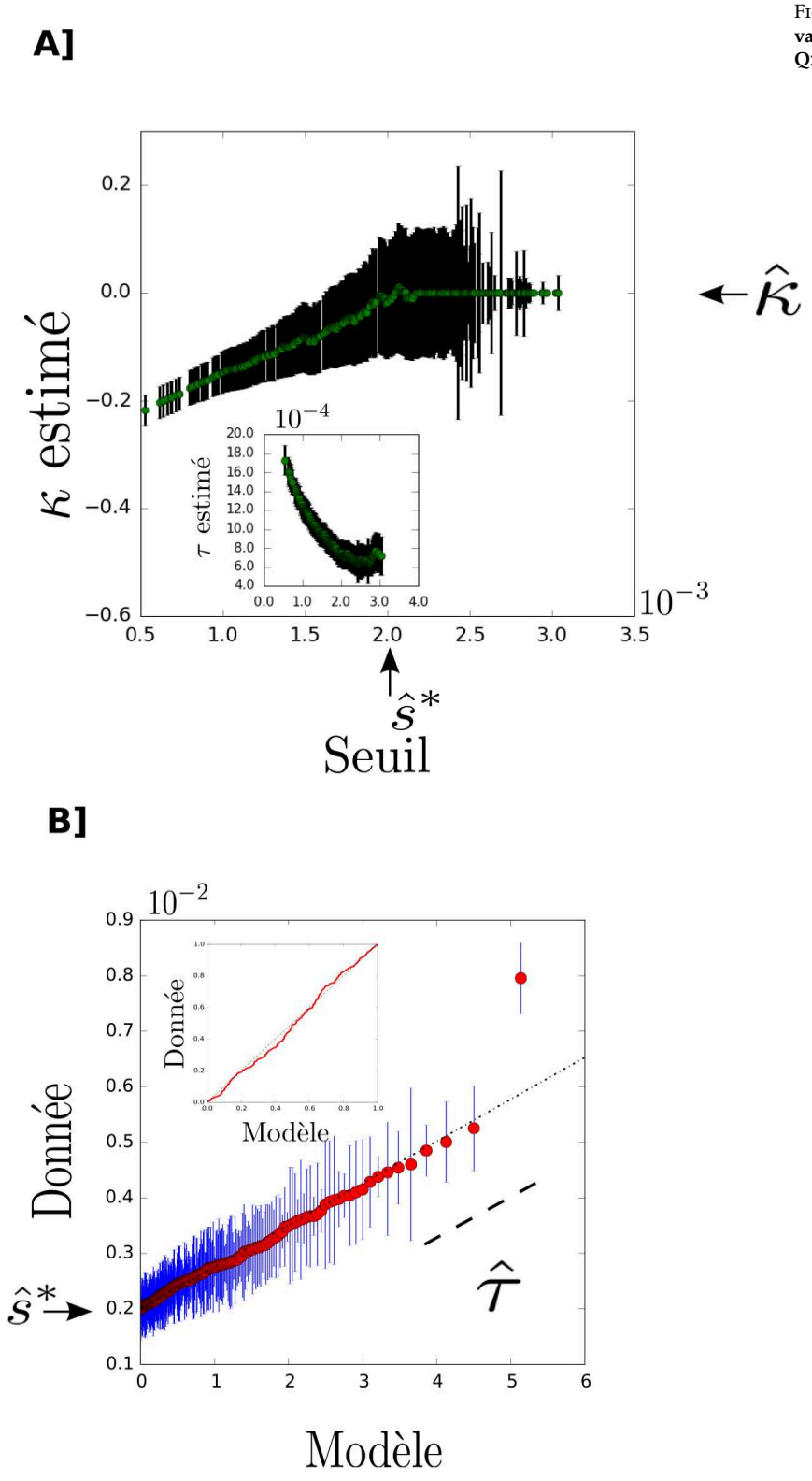


FIGURE 35: **Robustesse de l'analyse en valeurs extrêmes Frog3 contre PVP Q>30** $\frac{\Delta s}{s} < 0.3$

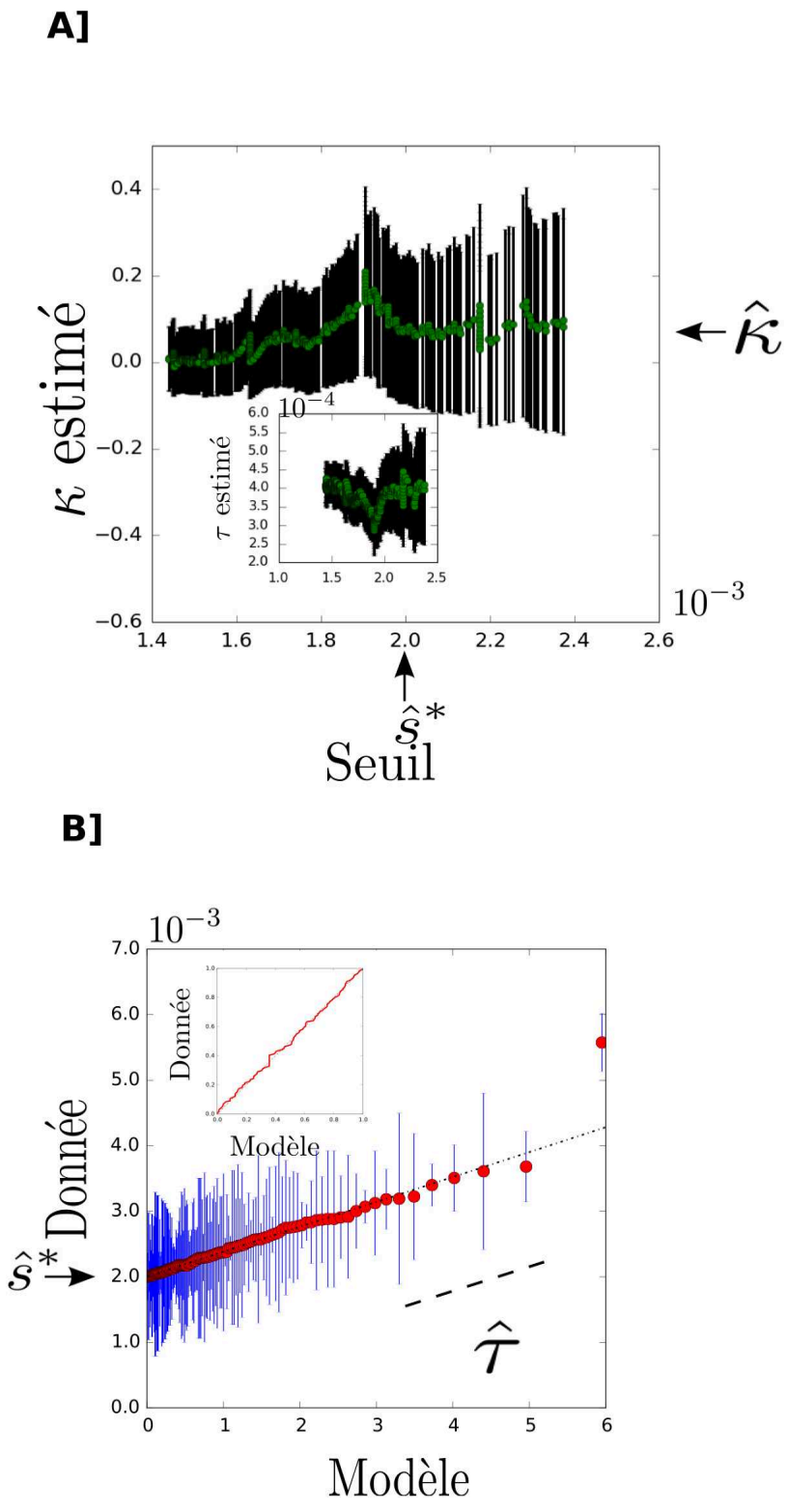


FIGURE 36: **Robustesse de l'analyse en valeurs extrêmes Frog3 contre PVP Q>30** $\Delta s < s_{noise}$

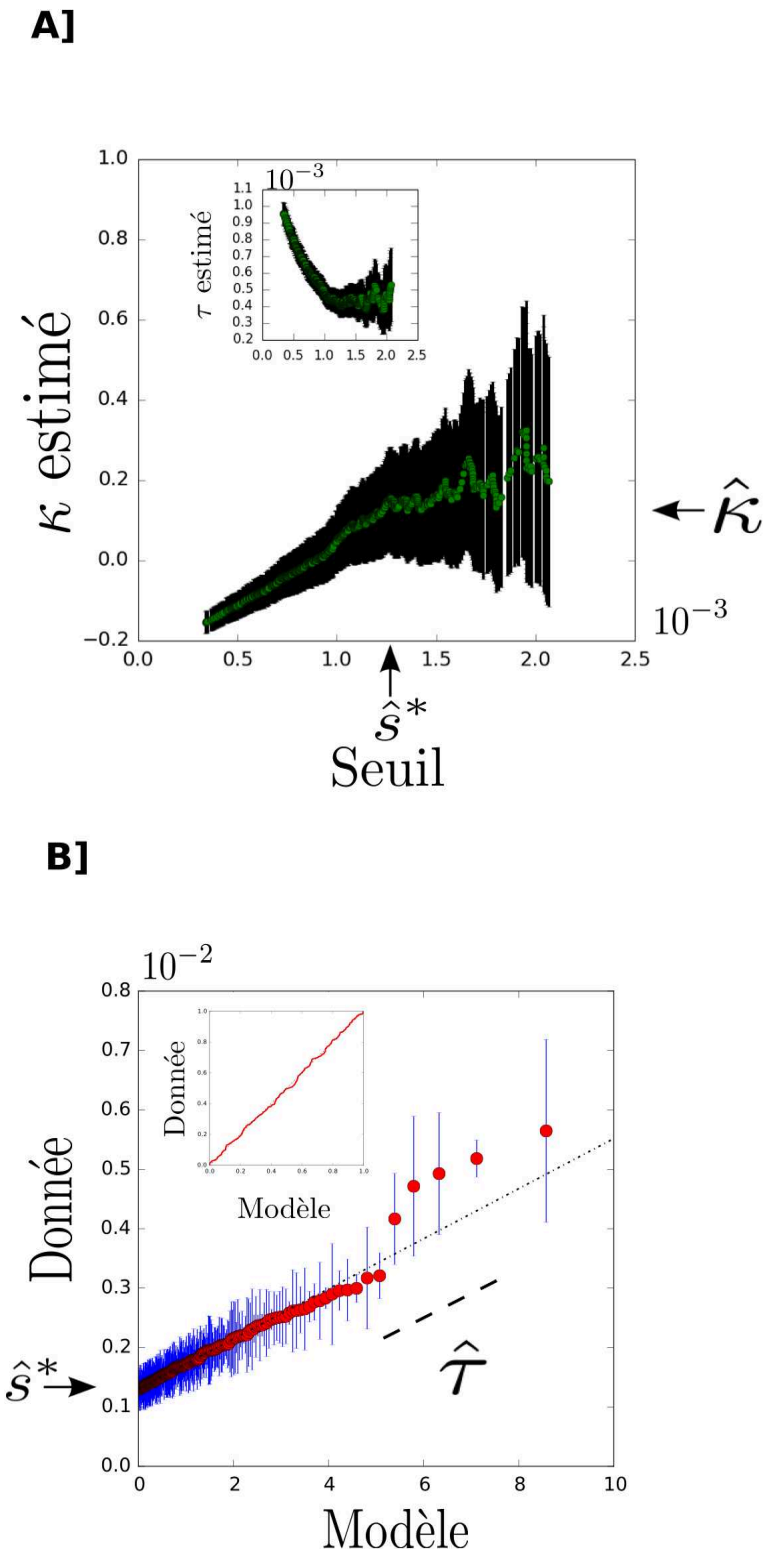


FIGURE 37: **Robustesse de l'analyse en valeurs extrêmes Frog3 contre PVP Q>o** $\frac{\Delta s}{s} < 0.3$

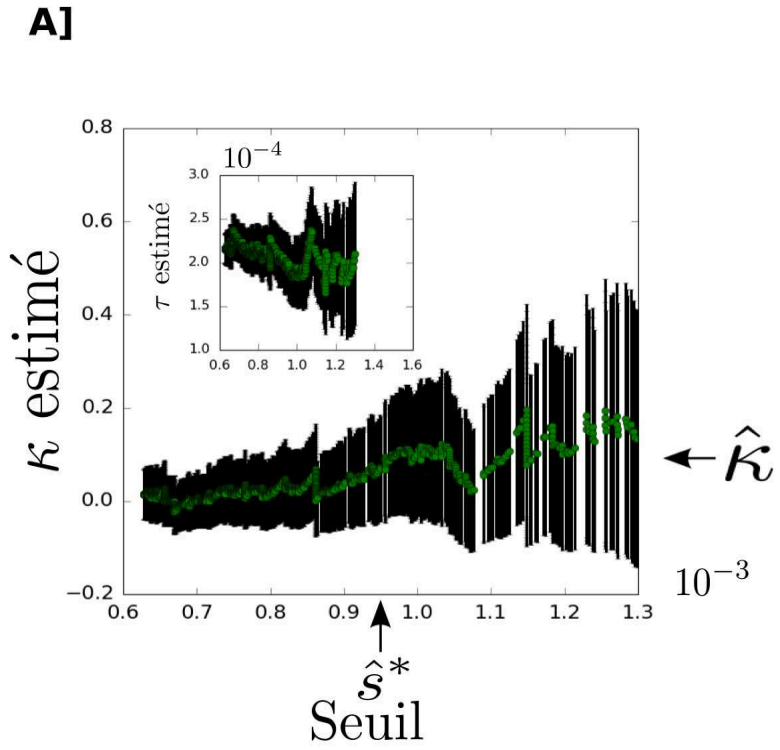
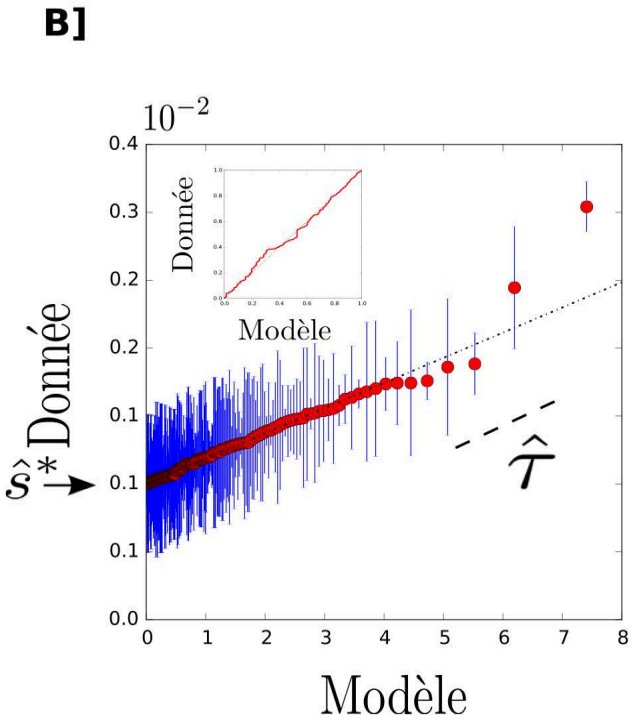


FIGURE 38: **Robustesse de l'analyse en valeurs extrêmes Frog3 contre PVP Q>0** $\Delta s < s_{noise}$



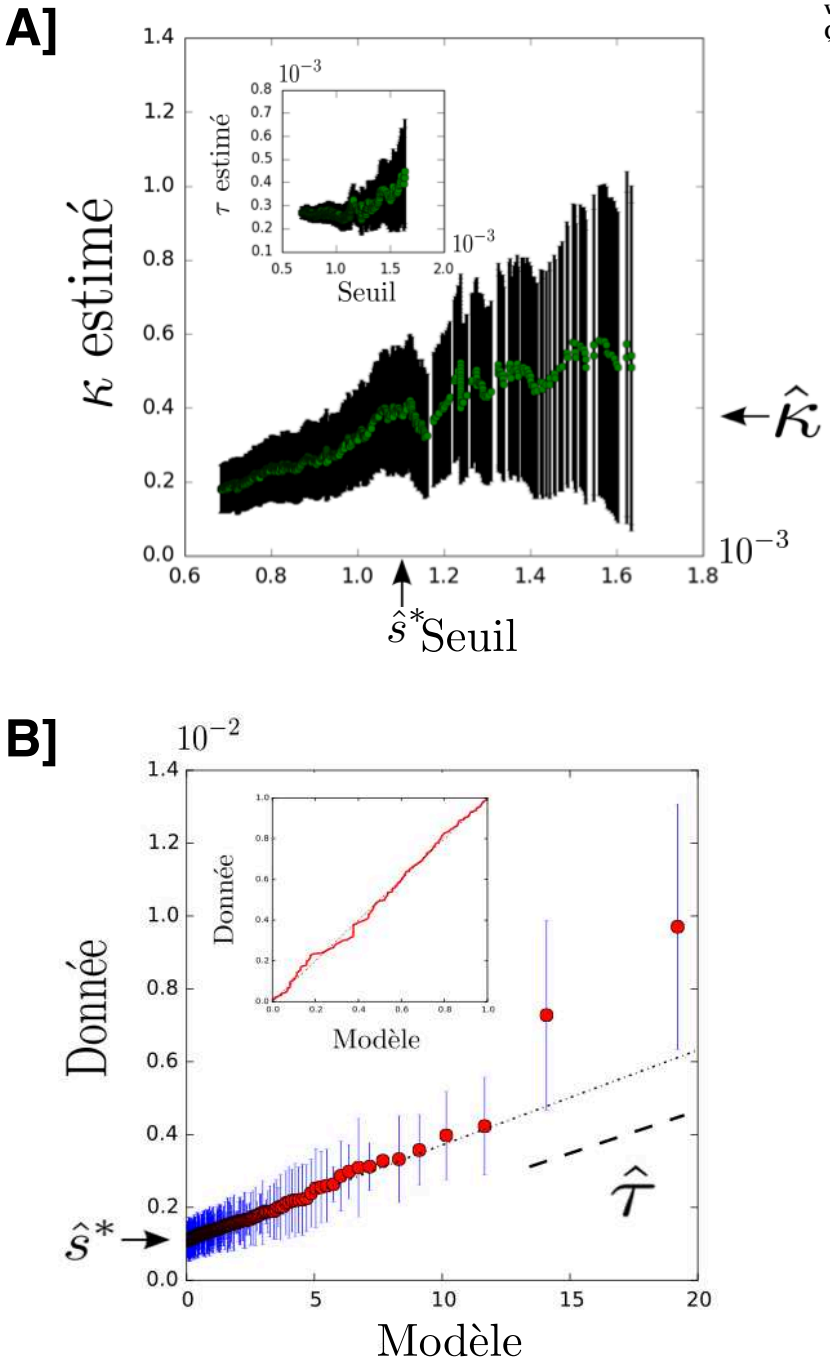


FIGURE 39: **Robustesse de l'analyse en valeurs extrêmes Frog3 contre PVP Q>o** $n > 10$

Valeurs extrêmes Frog3 contre noire

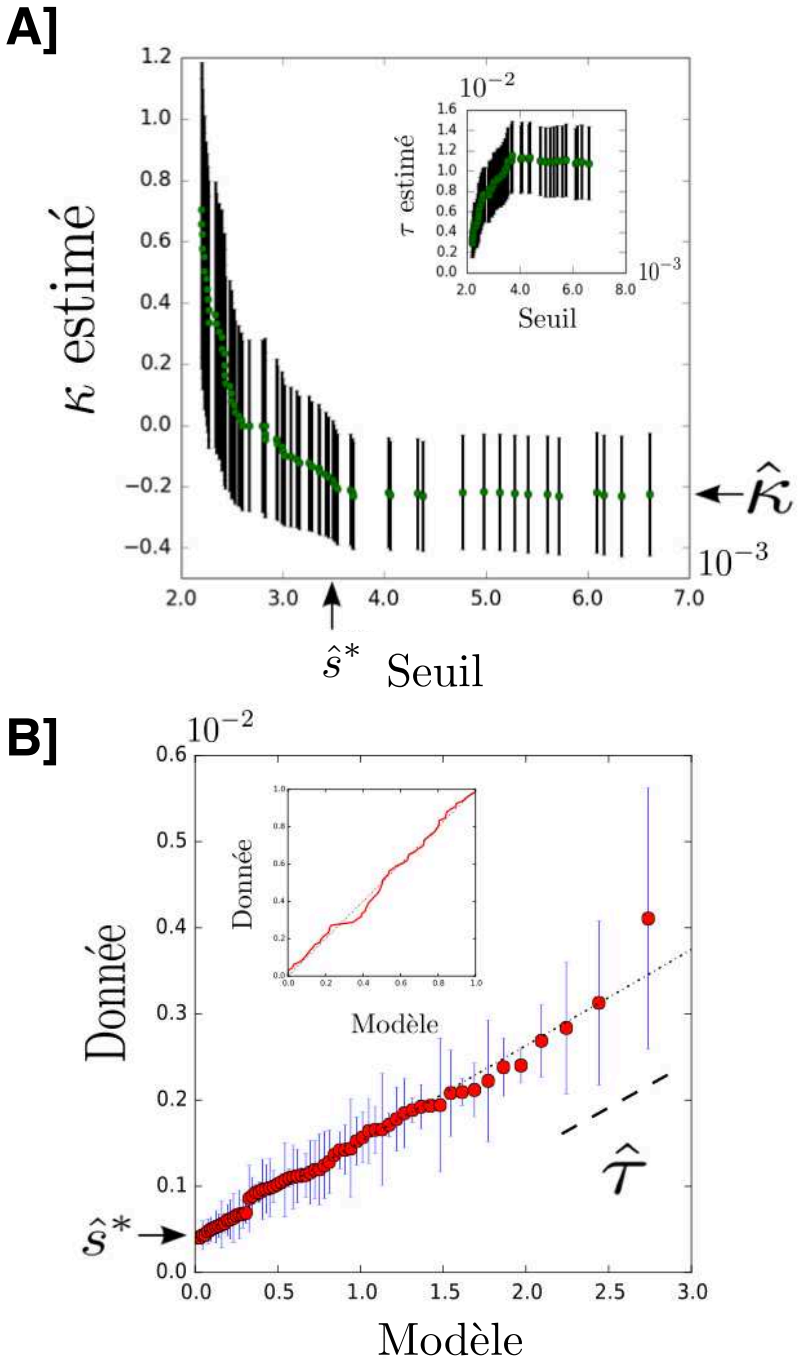
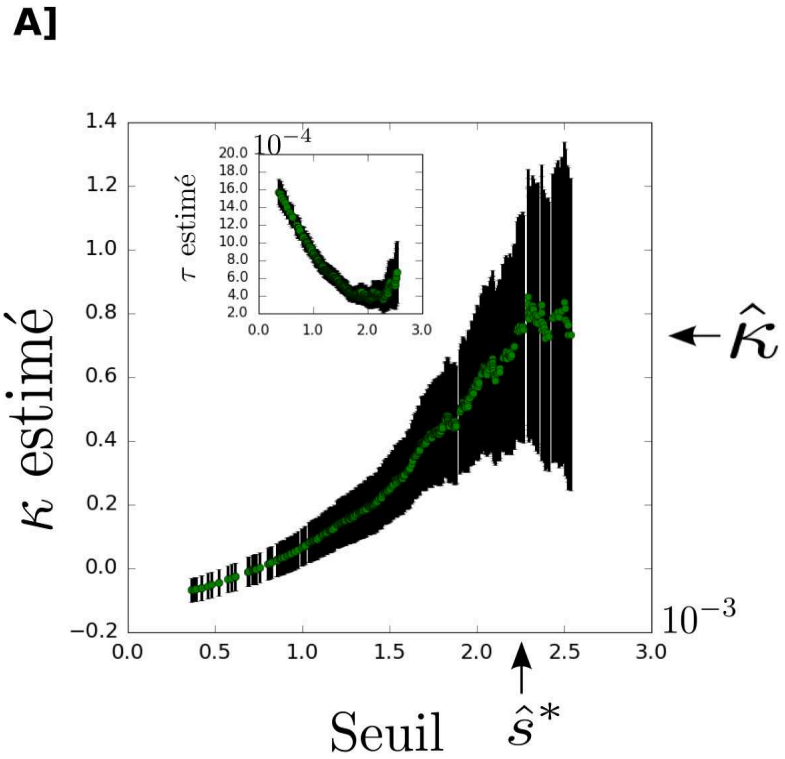


FIGURE 40: **Robustesse de l'analyse en valeurs extrêmes Frog3 contre noire Q>0** $n > 10$

Valeurs extrêmes Human2 contre noire



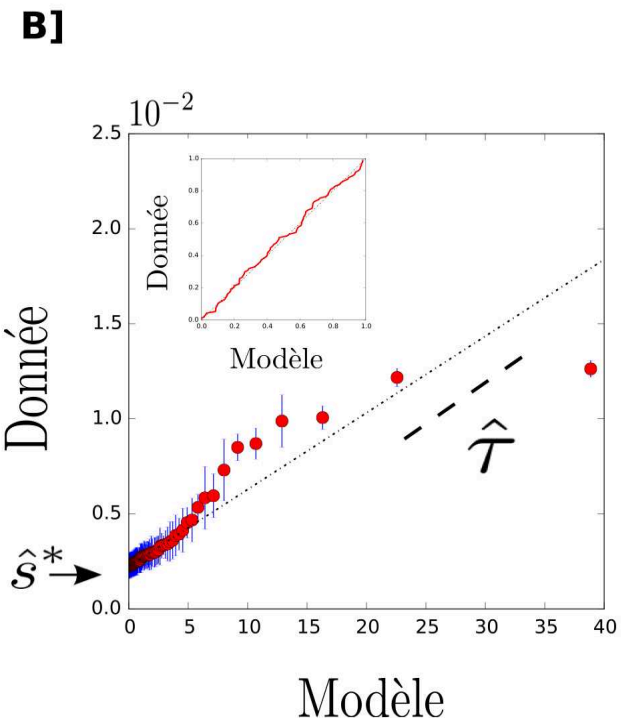


FIGURE 41: **Robustesse de l'analyse en valeurs extrêmes Human2 contre noire Q>30** $\frac{\Delta s}{s} < 0.3$

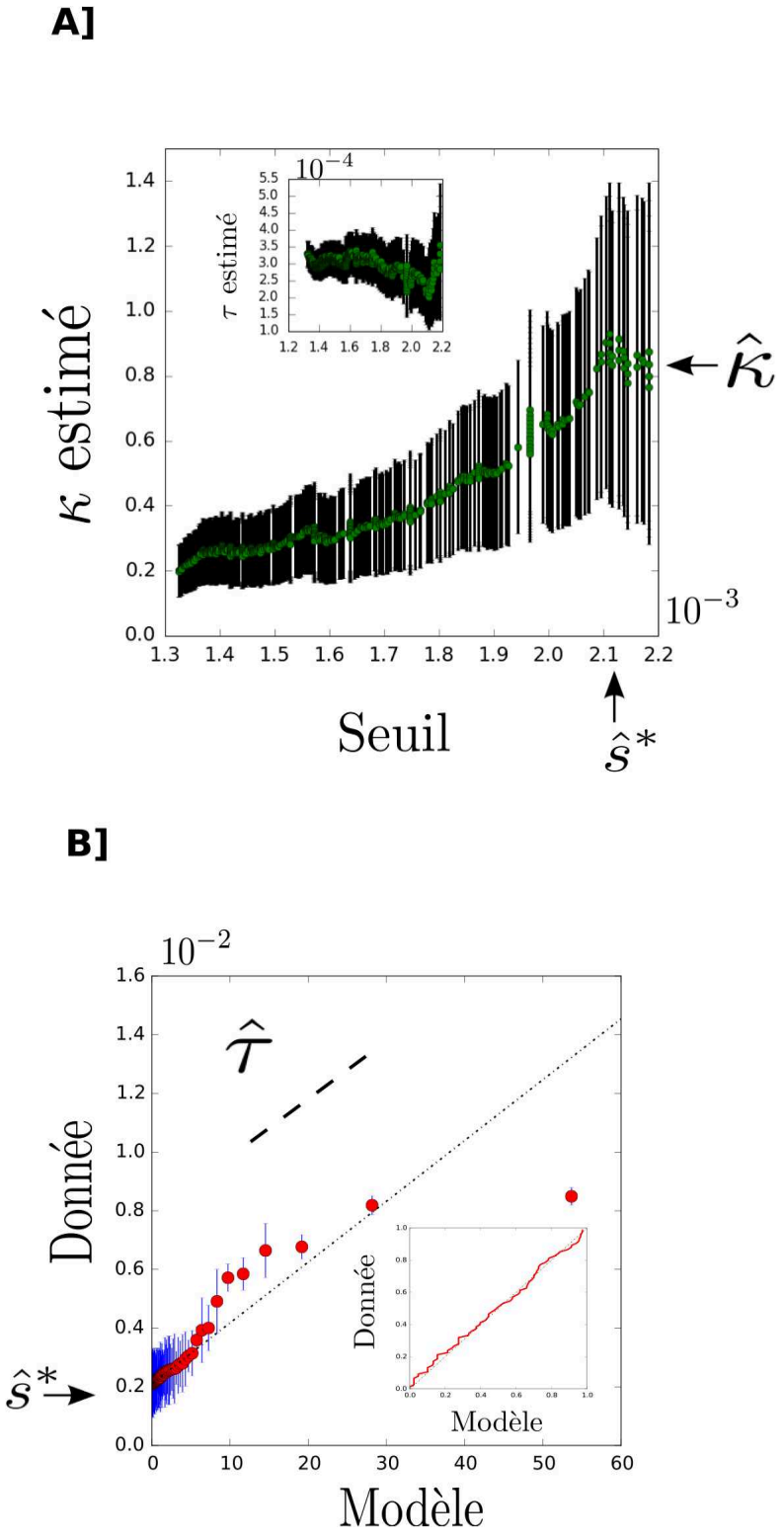


FIGURE 42: **Robustesse de l'analyse en valeurs extrêmes Human2 contre noire Q>30** $\Delta s < s_{noise}$

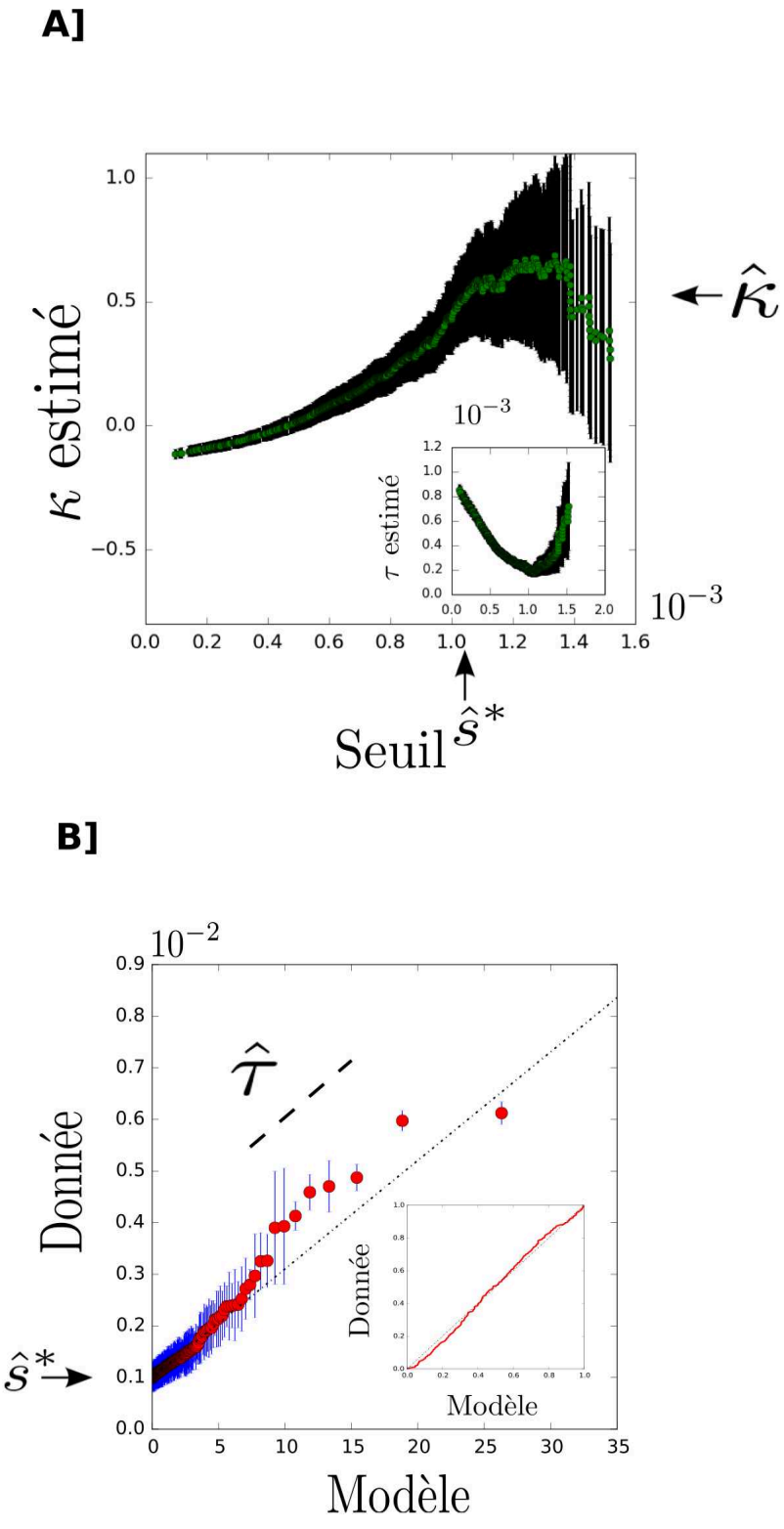


FIGURE 43: **Robustesse de l'analyse en valeurs extrêmes Human2 contre noire Q>0** $\frac{\Delta s}{s} < 0.3$

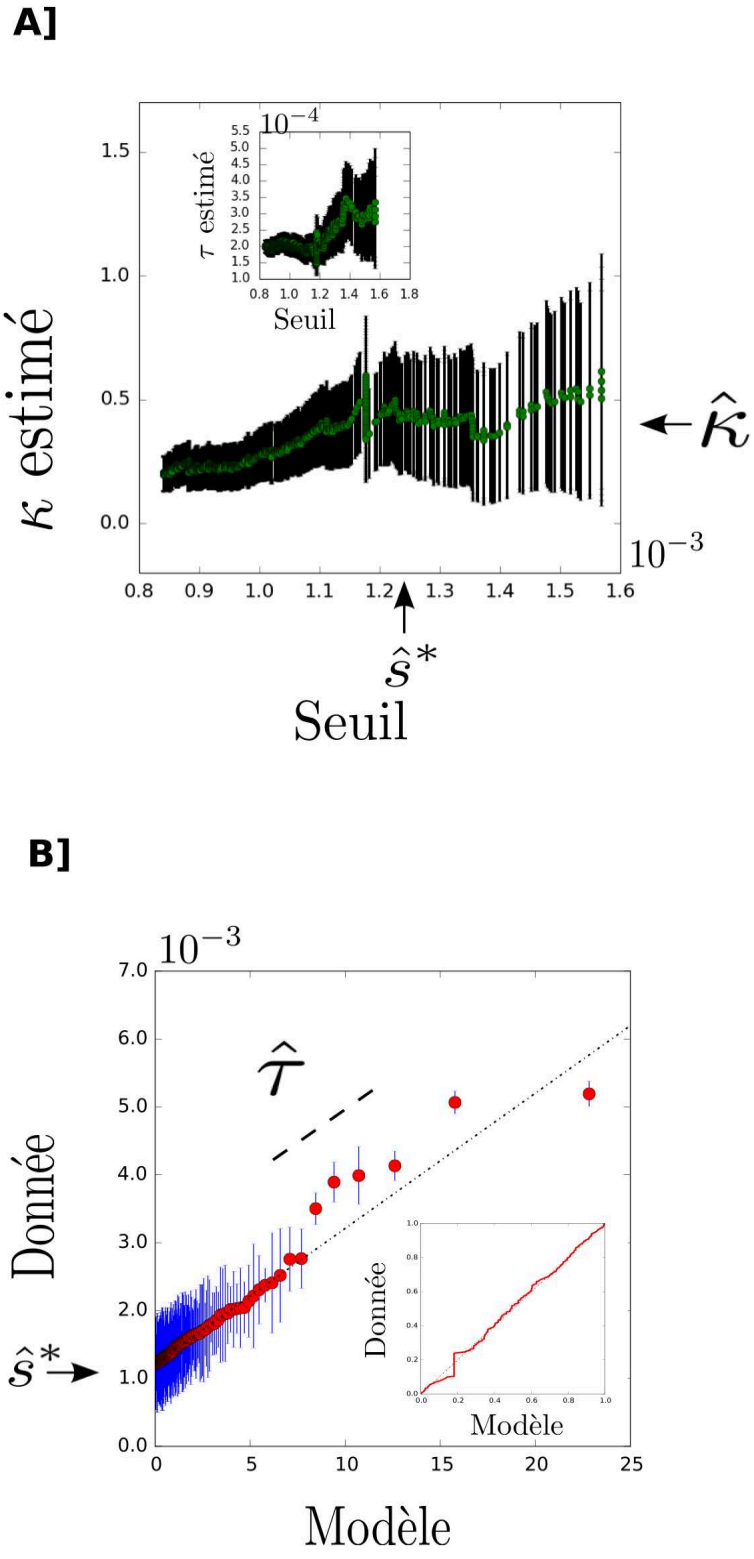


FIGURE 44: **Robustesse de l'analyse en valeurs extrêmes Human2 contre noire Q>0** $\Delta s < s_{noise}$

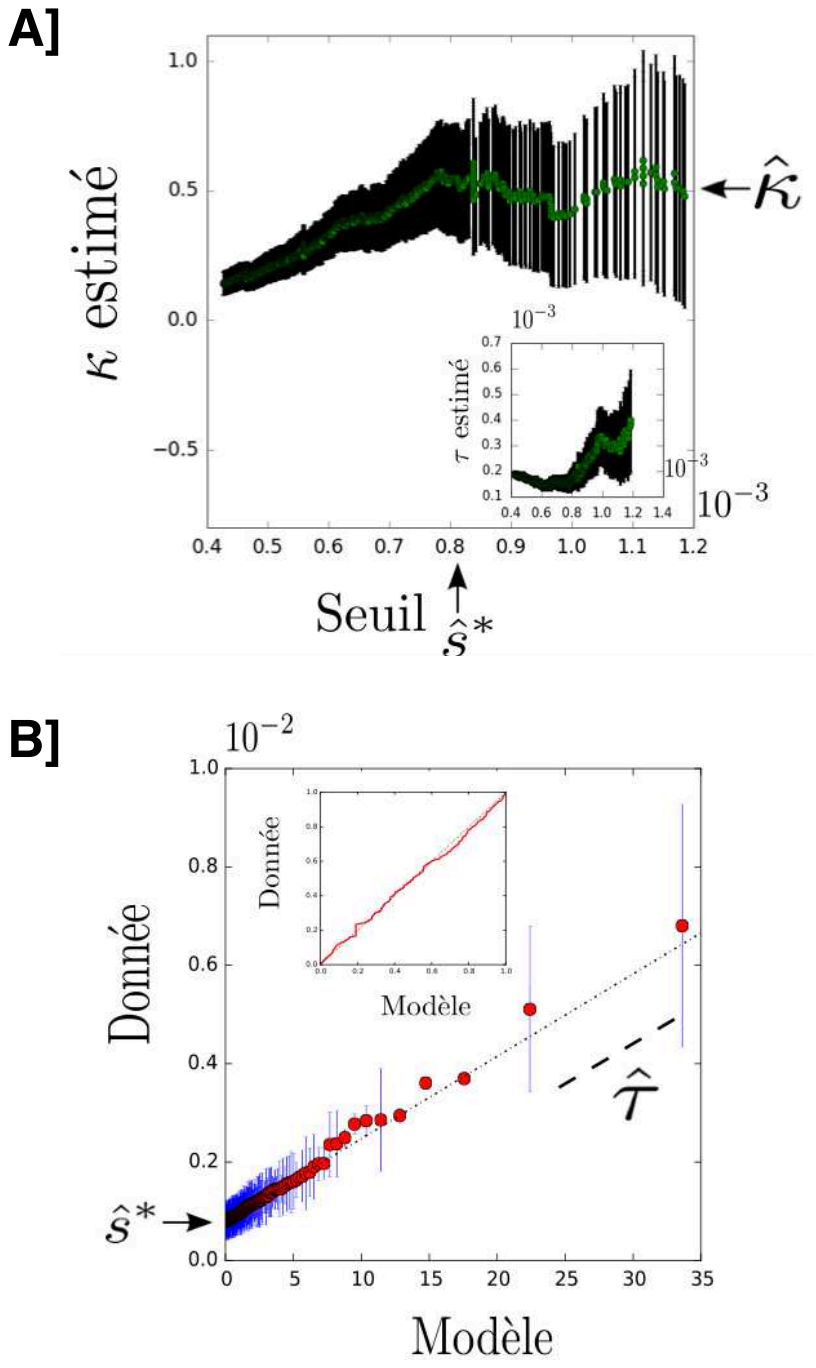
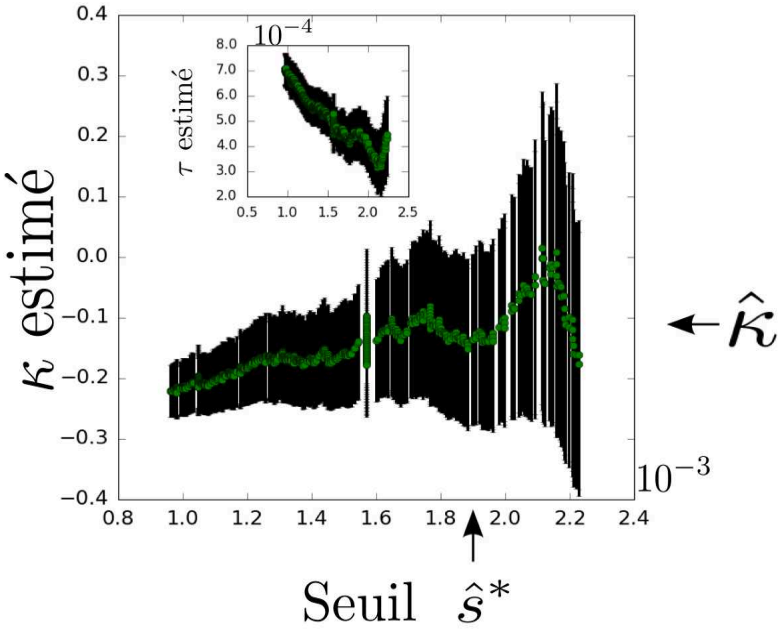


FIGURE 45: **Robustesse de l'analyse en valeurs extrêmes Human2 contre noire Q>0** $n > 10$

Valeurs extrêmes NoFramework contre noire

FIGURE 46: **Robustesse de l'analyse en valeurs extrêmes NoFramework contre noire Q>30** $\Delta s < s_{noise}$

A]



B]

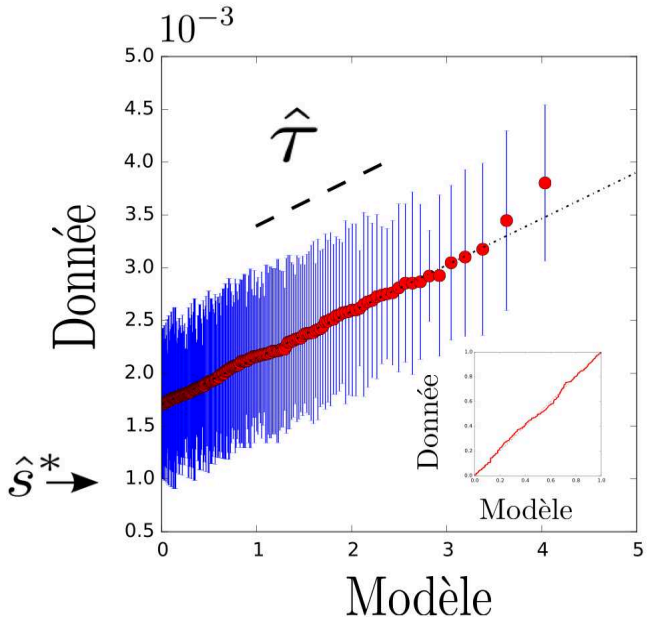
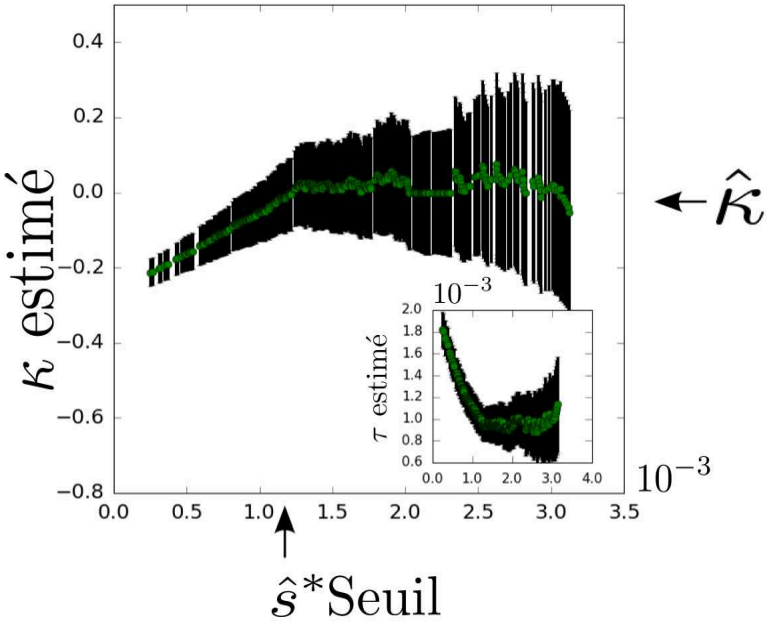
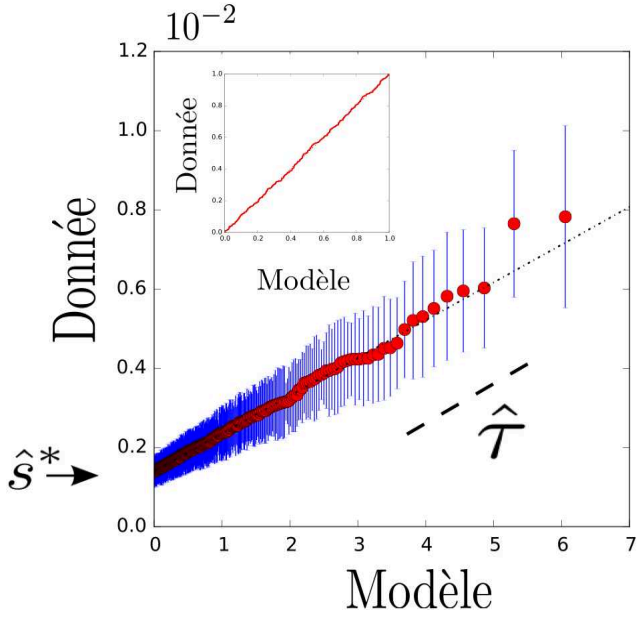


FIGURE 47: **Robustesse de l'analyse en valeurs extrêmes NoFramework contre noire Q>o** $\frac{\Delta s}{s} < 0.3$

A]



B]



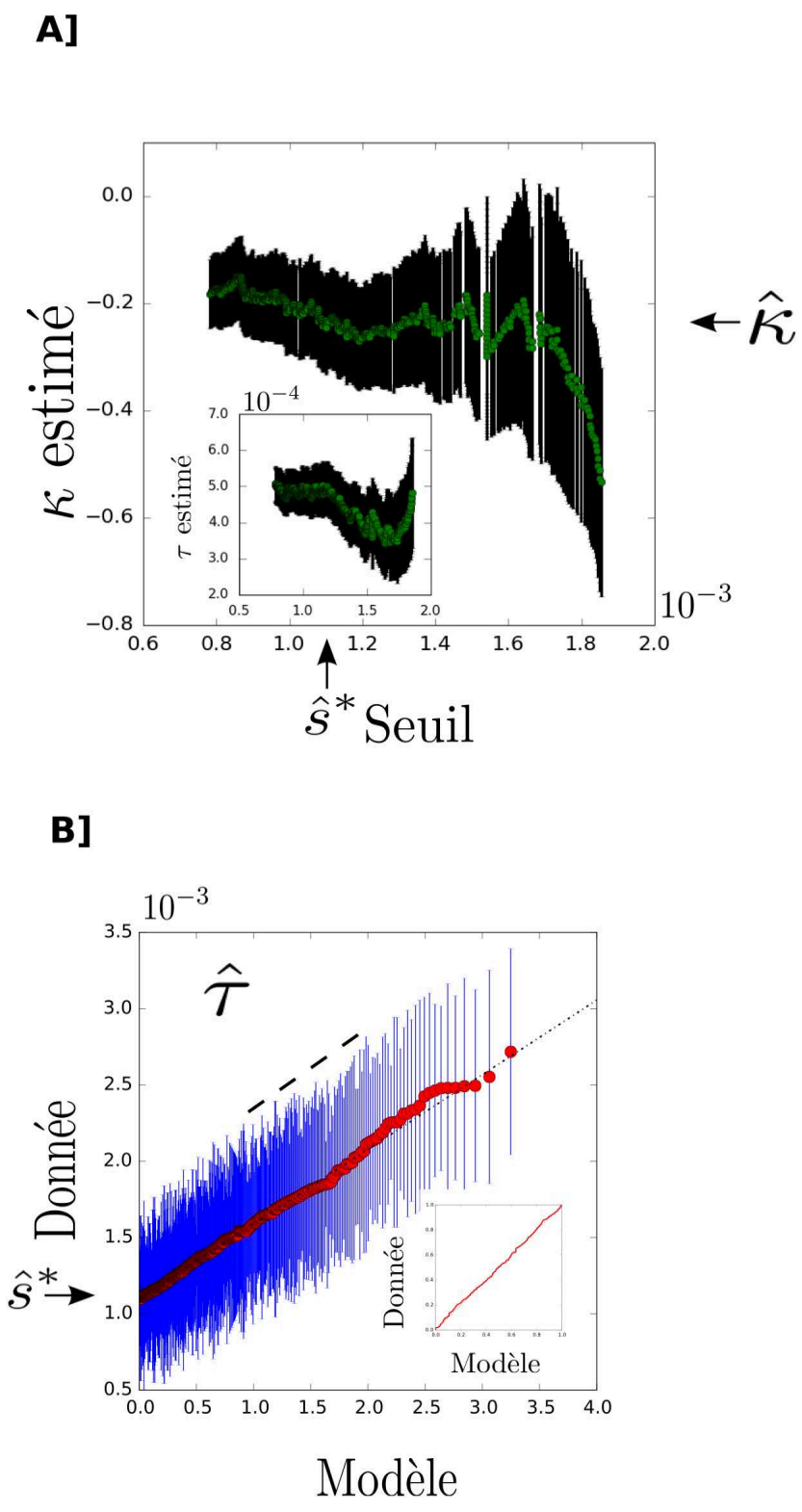


FIGURE 48: **Robustesse de l'analyse en valeurs extrêmes NoFramework contre noire Q>o** $\Delta s < s_{noise}$

Valeurs extrêmes Mix21bis contre noire

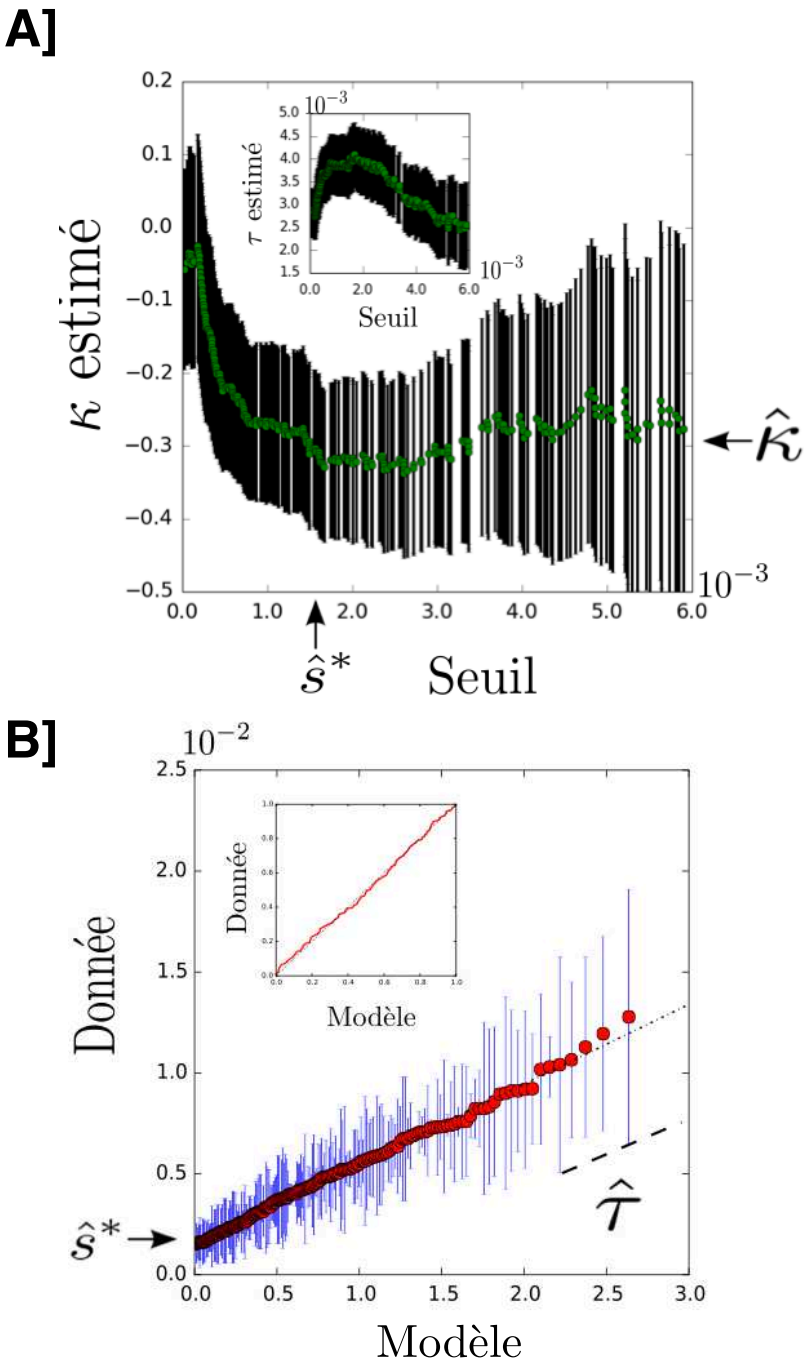
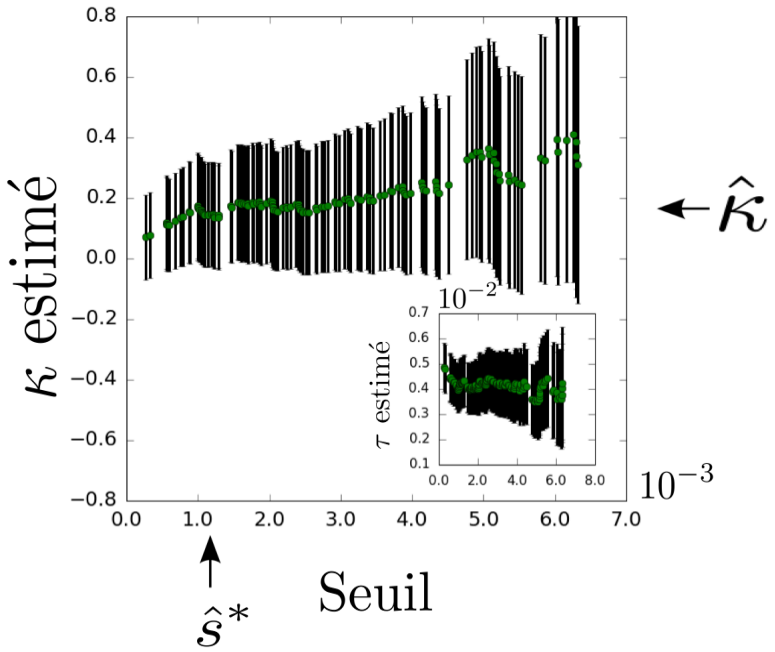


FIGURE 49: **Robustesse de l'analyse en valeurs extrêmes Mix21bis contre noire Q>0** $n > 10$

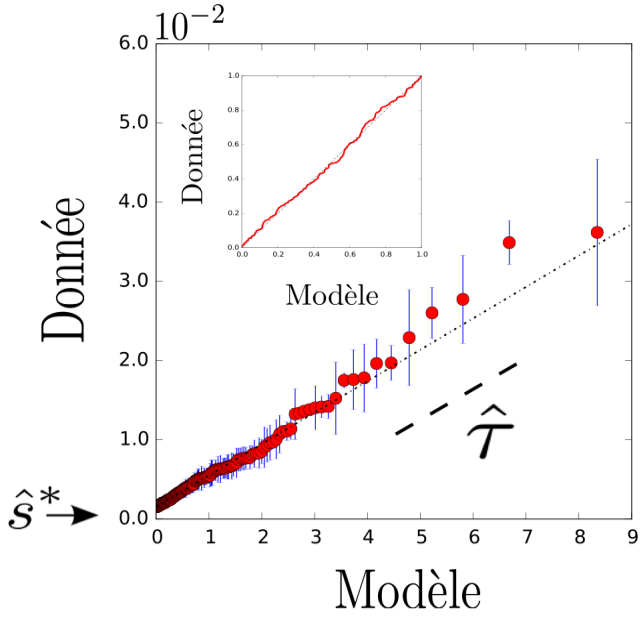
Valeurs extrêmes Mix24 contre noire

FIGURE 50: **Robustesse de l'analyse en valeurs extrêmes Mix24 contre noire Q>0** $\frac{\Delta s}{s} < 0.3$

A]



B]



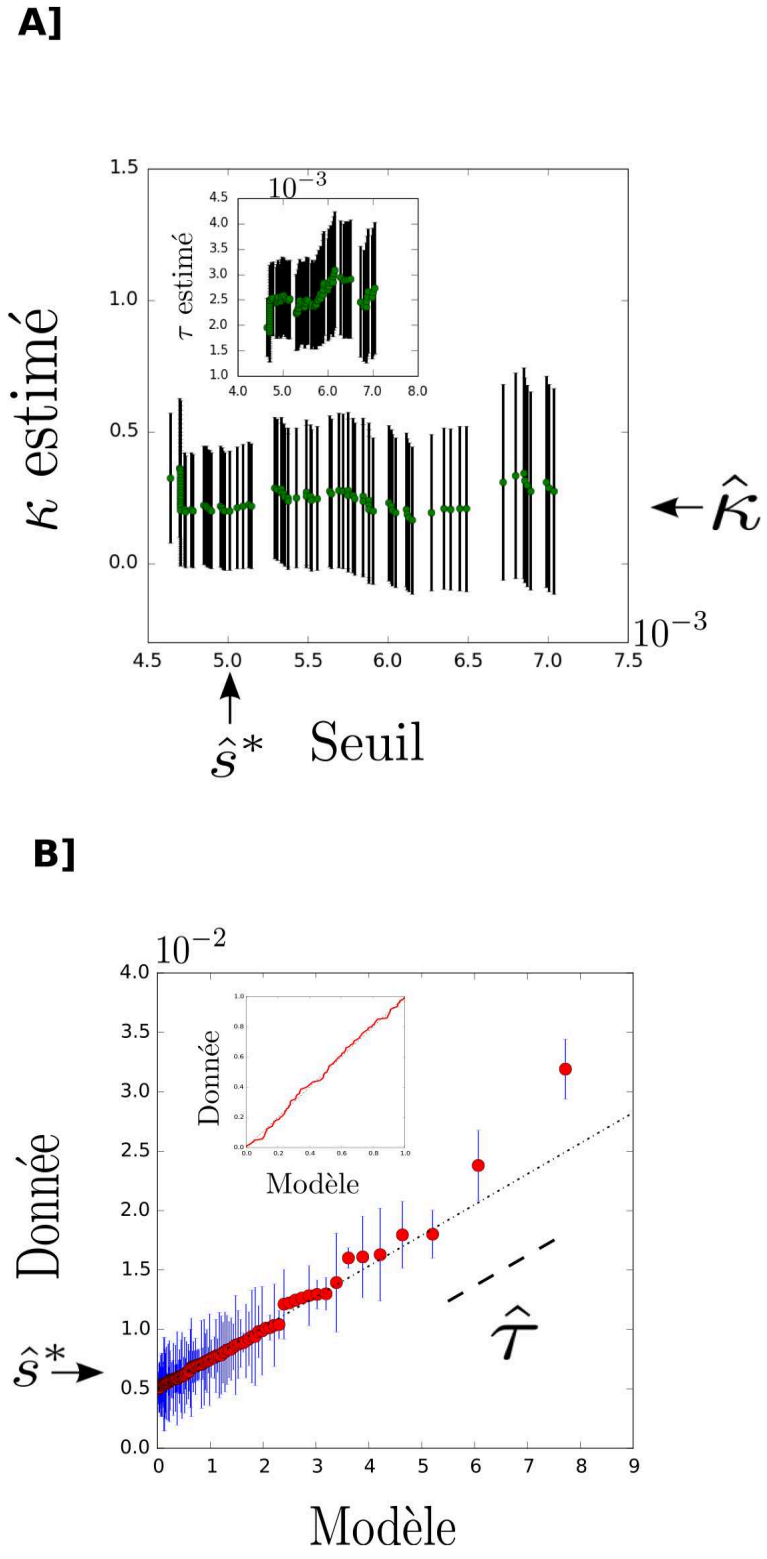


FIGURE 51: **Robustesse de l'analyse en valeurs extrêmes Mix24 contre noire Q>0** $\Delta s < s_{noise}$

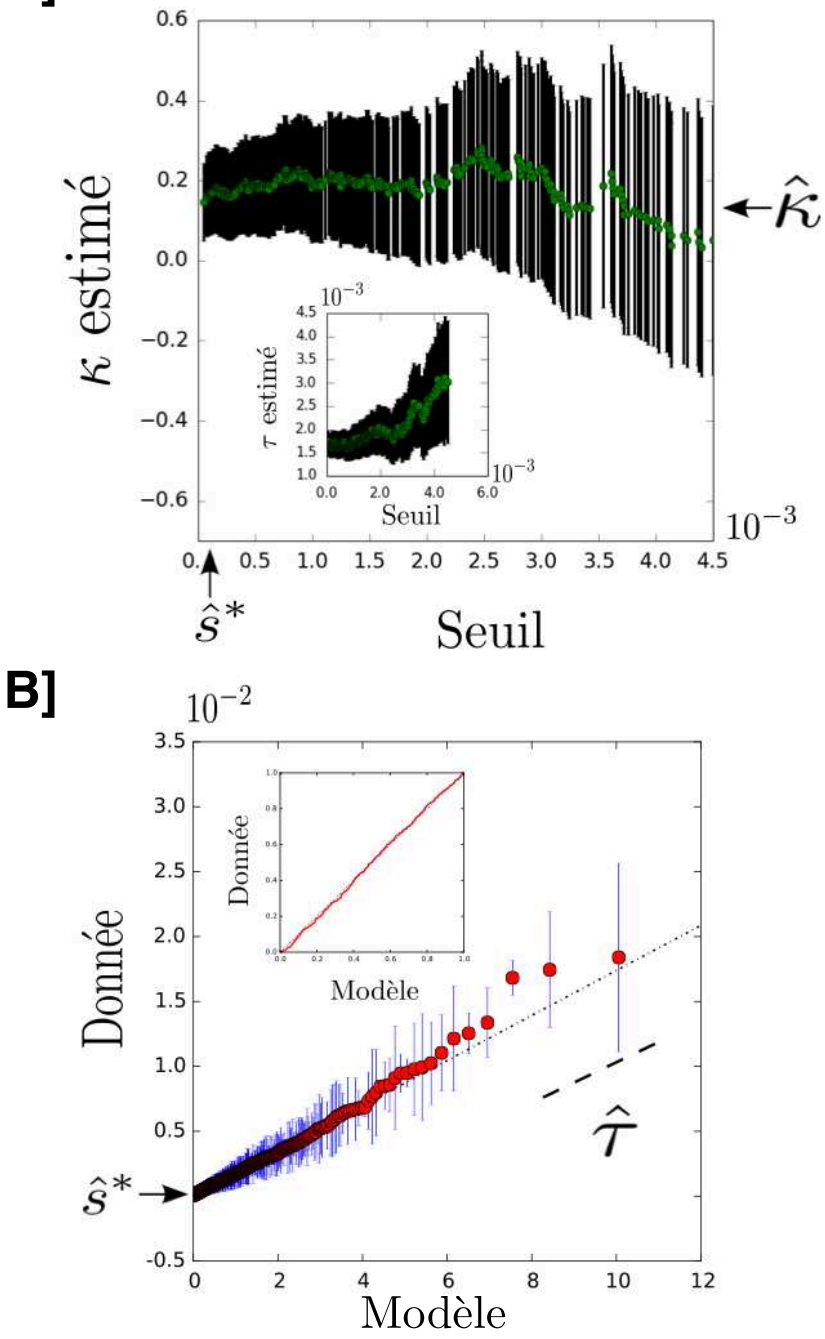


FIGURE 52: **Robustesse de l'analyse en valeurs extrêmes Mix24 contre noire Q>o** $n > 10$

Protocoles

Tous les produits chimiques ont été achetés chez Sigma-Aldrich (St Louis, MO), à moins qu'il ne soit précisé le contraire. De l'eau déionisée de résistivité 16MΩ.cm a été produite grâce à une résine d'échange d'ion (Aquadem(R) system of Veolia, Lyon, France). Un milieu de culture 2xTY est préparé en dissolvant 16g de tryptone yeast extract, 5g de NaCl (tryptone et yeast extract viennent de US-BIO distribué en France par Euromedex, Strasbourg, France) dans 1L d'eau déionisée et autoclavée pendant 15 min à 120°C.

Fonctionnalisation du PVP sur boites de Pétri et plaques 96 puits en polystyrene.

La fonctionnalisation du PVP c'est faite suivant le protocole décrit dans une précédente étude [1]. En résumé, du PVP à haute masse moléculaire (en moyenne MW 1300000) a été immobilisé sur des plaques 96 puits Greiner en polystyrene non traitées (Thermo Fisher Scientific, Waltham, MA) et des boites de Pétri Nunc en polystyrene non traitées de 3cm de diamètre (Thermo Fisher Scientific, Waltham, MA) en utilisant une technique de photo accrochage en UV profond grâce à de l'AziGrip4(Patent EP2236524, Adhesion promoter based on a functionalized macromolecule comprising photoreactive groups SuSoS AG, Dubendorf, Switzerland, www.susos. Com). Les plaques ont été tout d'abord traitées au plasma oxygéné pendant 2 min, incubées pendant 30 min dans une solution de AziGrip4 a 0.1 mg/mL, et lavées 3 fois avec de l'eau ultra pure. Un film de PVP a alors été déposé sur chaque plaque ou puit, grâce à une solution de 25mg/mL de PVP dans de l'éthanol. Après un séchage d'une nuit, les plaques ont été passées sous UV-C pendant 2 min. L'excès de PVP non attaché est enlevé par immersion des plaques dans de l'eau déionisée pendant une nuit et rinçage par de l'eau ultra pure. Cette procédure conduit à la formation d'un film dense de PVP attaché d'une manière covalente, et d'épaisseur approximative de 10 à 20 nm d'épaisseur.

1. Ananda Kumar Soshee et al. General in vitro method to analyze the interactions of synthetic polymers with human antibody repertoires. *Biomacromolecules*, 15:113–121, 2014

Cible ADN et immobilisation.

Les cibles ADN ont été préparées par auto-assemblage de la boucle ADN ayant à son bout 5', de la biotine. Les oligonucléotides ont été achetés chez Eurogentec (Anger, France). Ces oligonucléotides ont été purifiés par electrophorése sur gel polyacrylamide par le four-

nisseur et reçus sous forme lyophilisée. Les cibles ont été resuspendues dans de l'eau deionisée pour atteindre une concentration stock de 400μM. 20mg de billes magnétiques recouvertes de streptavidine (Dynabeads(R) M-280 Streptavidin de chez Invitrogen Life Technologies SAS, Saint Aubin, France) ont été préparés en suivant le protocole de Dynabeads(R). 10μL de boucles d'ADN du stock ont été mixés avec 20mg de Dynabeads(R) lavés et incubés pendant 10 min à température ambiante sous agitations légères. Les billes magnétiques alors couvertes des cibles ADN sont séparées du milieu à l'aide d'un aimant et lavées 3 fois avec une solution tampon contenant 5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA et 1M NaCl.

Sélection et amplification par phage display.

La production de phages et la sélection par phage display de nos banques a été faite en suivant en grande partie le protocole standard venant de Source BioScience (Cambridge, U.K. ; http ://lifesciences.sourcebioscience.com/media/143421/tomlinsonij.pdf) et de nos travaux précédents[2] [3] avec quelques modifications qui seront explicitées ci-dessous.

2. Ananda Kumar Soshee et al. General in vitro method to analyze the interactions of synthetic polymers with human antibody repertoires. *Biomacromolecules*, 15:113–121, 2014

3. Purvi Jain et al. Selection of arginine-rich anti-gold antibodies engineered for plasmonic colloid self-assembly. *J. Phys. Chem. C*, 118:14502–14510, 2014

Production de phages

Des phages à une concentration finale de 10^{11} phages/ml sont préparés par superinfection de TG1 E. Colie (Source Bioscience, Cambridge, U.K.) en phase exponentielle avec des phages Helper M13 KO7 (GE Healthcare, Pittsburgh, PA). Les E. Colies contenant le plasmide codant pour la fusion pIII/anticorps de la protéine de surface du phage sont laissées en pousse d'un D.O. 0.06 à une D.O. 0.6 dans du milieu 2xTY auquel on a ajouté 100μg/mL d'ampicilline et 1%masse/volume de glucose. A une D.O. de 0.6, la banque d'E. Colie est superinfectée par le phage helper pendant 30 min à 37°C. Après superinfection les bactéries sont centrifugées à 3300g pendant 10 min et le surnageant est enlevé. Les bactéries sont alors resuspendues dans un nouveau milieu 2xTy sans glucose mais avec ampicilline et kanamycin. Les Tg1 infectées poussent alors pendant 7h à 30°C. Elles sont ensuite centrifugées pendant 10 min à 10 000g. On ne garde alors que le surnageant contenant les phages qui sera gardé à 4°C pour être utilisé le jour suivant.

Phage display

Lors de l'expérience de phage display, le surnageant de phages contenant notre banques est ajusté pour baigné dans une solution de 1xPBS sans NaCl. Les phages sont alors d'abord présentés soit à des billes magnétiques sans cibles ADN soit à du polystyrene non traité, pour sélection négative. Pour la sélection contre les cibles ADN, des tubes Eppendorf DNA LoBind (Eppendorf AG, Hamburg, Germany) sont utilisés. Les phages sont incubés à température ambiante 1h sans agitations puis 30 min sur une roue de mélange. Les phages restants (ceux qui ne se sont pas liés lors de cette étape de sélec-

tion négative) sont alors présentés à leur cibles respectives (boucle ADN ou PVP). Dans le cas des cibles ADN, 50μL de billes (environ 10^{13} cibles potentielles) sont utilisés. Lors de cette étape un excès de boucle d'ADN (10^{14}) est présenté aux billes. Le protocole pour le lavage des cibles est celui de la notice commerciale fournie par Dynabeads(R). La selection des anticorps est alors faite contre les billes recouvertes d'ADN ou le PVP, pendant 1h30 sur un plan mélangeur. Il s'en suit 10 étapes de lavages avec un mélange de 1xPBS + 0.1% Tween20. Puis les phages encore accrochés aux cibles sont décrochés grace à 1mL de solution de triethylamine à 100mM pendant 20 min. Cette solution est ensuite neutralisée avec 500 µL of Tris/HCl solution tampon (1 M, pH 7.4). Les phages ainsi élués servent alors à l'infection d'un excès de E. Colie en phase exponentielle (14 mL de culture dans 2xTY à O.D.600 nm = 0.6), qui sera titré et servira à la préparation de phages pour une prochaine étape de sélection. Les TG1 ainsi infectés sont déposées sur une gélose de 2xTy + ampicilline et poussent toute la nuit à 37°C pour amplification. A partir de cette culture un stock glycerol est préparé pour stockage à -80°C.

Préparation pour séquençage

La banque de phagemides est purifiée avec un kit Midiprep de Macherey-Nagel (Hoerdt, France). La technologie Illumina MiSeq avec kit de chimie v3 est utilisée pour le séquençage par la société Eurofins Genomics (Ebersberg, Germany).

Bibliographie

[1] Ananda Kumar Soshee et al. General in vitro method to analyze the interactions of synthetic polymers with human antibody repertoires. *Biomacromolecules*, 15 :113–121, 2014.

[2] Purvi Jain et al. Selection of arginine-rich anti-gold antibodies engineered for plasmonic colloid self-assembly. *J. Phys. Chem. C*, 118 :14502–14510, 2014.

Mots-clefs : evolution, physique statistique, anticorps.
L'évolution par sélection naturelle se compose d'une succession de trois étapes : mutations, sélection et prolifération. Nous nous intéressons à la description et à la caractérisation du résultat d'une étape de sélection dans une population composée de nombreux variants. Après sélection, cette population va être dominée par les quelques meilleurs variants, ceux qui ont la plus grande capacité à être sélectionnés, ou plus grande « sélectivité ». Nous posons la question suivante : comment est distribuée la sélectivité des meilleurs variants dans la population ? La théorie des valeurs extrêmes, qui caractérise les queues extrêmes des distributions de probabilités en terme de 3 classes d'universalités, a été proposée pour répondre à cette question. Pour tester cette proposition et identifier les classes d'universalités rencontrées dans ce genre de problème, nous avons procédé à une sélection quantitative de banques composées de 10^5 variants d'anticorps grâce à la technique du phage display. Les données obtenues par séquençage à haut débit du résultat de la sélection de nos banques nous permettent d'ajuster la distribution de sélectivités obtenue sur plus de deux décades.

Keywords : evolution, statistical physics, antibodies.
Evolution by natural selection involves the succession of three steps : mutations, selection and proliferation. We are interested in describing and characterizing the result of selection over a population of many variants. After selection, this population will be dominated by the few best variants, with highest propensity to be selected, or highest "selectivity". We ask the following question : how is the selectivity of the best variants distributed in the population ? Extreme value theory, which characterizes the extreme tail of probability distributions in terms of a few universality class, has been proposed to describe it. To test this proposition and identify the relevant universality class, we performed quantitative in vitro experimental selections of libraries of $> 10^5$ antibodies using the technique of phage display. Data obtained by high-throughput sequencing allows us to fit the selectivity distribution over more than two decades. In most experiments, the results show a striking power law for the selectivity distribution of the top antibodies, consistent with extreme value theory.