

RAPPORT DE L'ÉPREUVE ÉCRITE DE BIOLOGIE

Coefficients (en pourcentage du total d'admission) :

ENS Paris-Saclay : 12,3%

ENS de Lyon : Option biologie : 13,2%, Option ST : 6,6%

ENS : Option biologie : 4,8%, Option ST : 2,7%

ENPC : 5,0%

Membres du jury :

Guillaume Barthole ; Lucie Bastin-Héline ; Camille Berthelot ; Alain Bessis, Eloïse De Tredern ; Gersende Lepère ; Benoît Maffei ; Anne Charlotte Marsollier, Déborah Prévôt, Stéphane Réty

L'épreuve de biologie de la session 2021, d'une durée totale de 6h, proposait aux candidates et aux candidats¹ de réfléchir autour du génome humain. Le sujet de synthèse (partie A) d'une durée conseillée de 2h traitait de la variabilité génétique dans le génome humain et l'étude de documents traitait des traces de l'histoire des humains dans le génome.

Le jury salue la très grande qualité des copies qui ont su faire preuve de synthèse et de rigueur dans leur appréhension des différentes thématiques.

SUJET DE SYNTHÈSE :

La variabilité génétique dans la population humaine

L'épreuve de synthèse permet d'évaluer les capacités de **communication, de rédaction** et d'**argumentation scientifique**. Le sujet proposé était proche des attendus du programme mais demandait une compréhension solide et approfondie des notions sous-jacentes pour en faire une synthèse organisée. Le sujet couvrait l'étude de la variabilité génétique chez les humains depuis son origine jusqu'à ses conséquences fonctionnelles et évolutives, et s'appuyait sur des notions à la fois de biologie moléculaire et de biologie des organismes. Ce sujet nécessitait donc une sélection rigoureuse des informations pertinentes à travers les connaissances des candidats.

1. Attendus et contenu

Délimitation du sujet

Une part essentielle du sujet s'appuyait sur les parties suivantes du programme : I-A-2 Les macromolécules ; III-A Les populations et leur dynamisme ; IV-A Génome des Eucaryotes et

¹ Ce rapport s'adresse à toutes et tous mais afin d'en alléger l'écriture, nous emploierons par la suite le masculin.

L'expression du génome : la transcription et son contrôle ; IV-B Réplication de l'information génétique et mitose; IV-C La diversification des génomes et IV-D Les mécanismes de l'évolution.

Trois aspects essentiels devaient être développés : (i) La nature et les origines de la variabilité génétique entre individus, (ii) sa transmission et son brassage et (iii) ses conséquences aux niveaux fonctionnel et populationnel. Sans forcément constituer un plan ou une progression, ces trois axes de lecture permettent d'aborder l'ensemble du sujet.

Le sujet était restreint à la population humaine, soit une espèce eucaryote à reproduction sexuée exclusive : les mécanismes moléculaires et cellulaires propres aux procaryotes étaient hors-sujet, tout comme le détail de certains exemples classiques portant sur la dynamique des génomes bactériens tels que l'expérience de Luria-Delbrück. La description détaillée des mécanismes moléculaires de la fourche de réplication, de la transcription et de la traduction n'étaient pas attendus, puisqu'ils ne sont pas directement liés à la variabilité génétique.

Notions attendues

La notion de mutation génétique comme source unique de la variabilité était centrale au sujet, et a été abordée par la majorité des candidats, avec des niveaux de précision variables. La structure de l'ADN, la nature des variations génétiques observées entre individus et leurs origines liées au fonctionnement cellulaire et au hasard ont souvent été bien traitées. Cependant, le jury a noté que seuls de rares candidats ont su mettre en contexte ces informations en les replaçant chez l'être humain – par exemple la taille du génome humain ou son nombre de chromosomes, informations parfois mentionnées mais souvent avec des valeurs incorrectes. Peu de candidats font la distinction entre lésion (modification de la structure de l'ADN), mutation (modification définitive et transmissible de la séquence nucléotidique, qui est le résultat d'un problème de réparation d'une lésion stabilisée dans le génome à la réplication suivante) et allèle (une des variations de séquence dans la population qui résulte d'une ou plusieurs mutations dans un allèle pré-existant). Les copies ayant fait preuve de rigueur sur ces aspects ont été valorisées.

L'importance de la réplication, de la méiose et de la reproduction sexuée dans la transmission et le brassage de la variabilité génétique chez l'homme a bien été identifiée par les candidats comme un aspect clé du sujet, et a été abordée dans la majorité des copies. Les brassages intra- et inter-chromosomiques sont presque toujours mentionnés, mais le jury a été surpris par le nombre d'erreurs et d'incompréhensions sur ces mécanismes fondamentaux du fonctionnement cellulaire. Seule une minorité de candidats a su faire un schéma du déroulement de la méiose sans erreur. Le génome mitochondrial et sa transmission, par ailleurs, n'ont pratiquement jamais été mentionnés, tout comme les chromosomes sexuels, absents de la quasi-totalité des copies. Le jury a également noté que de nombreux candidats ont entendu parler des transferts horizontaux, mais ceux-ci sont généralement mal compris (confusions avec le microbiote, les éléments transposables, etc.).

Les conséquences phénotypiques, populationnelles et évolutives de la variabilité génétique constituaient le troisième axe majeur du sujet et ont également été traitées, à des degrés divers, par la majorité des candidats. Les conséquences des variants génétiques codants sur les protéines et les

phénotypes sont généralement abordées, mais l'importance des variations dans l'ADN non-codant – la grande majorité du génome – est plus rarement mentionnée, même si un développement d'envergure n'était pas attendu. Les notions de génétique classique comme l'homo- et l'hétérozygotie ou la dominance et la récessivité des allèles sont rarement introduites, alors qu'elles sont fondamentales pour des notions plus complexes, et souvent abordées par les candidats, comme l'équilibre de Hardy-Weinberg. Enfin, la sélection naturelle et la dérive génétique étaient deux points importants à traiter dans le cadre de ce sujet. Ces notions de génétique des populations semblent souvent mal comprises par les candidats : parmi les erreurs fréquentes, le jury a noté que la sélection est fréquemment présentée comme une source de diversité génétique (alors qu'elle est faible chez l'Homme et tend à diminuer la variabilité, sauf dans certains cas particuliers), et que la dérive est souvent présentée comme absente chez l'Homme (alors qu'il s'agit de la principale force évolutive dans notre espèce).

En conclusion, le jury félicite les candidats qui dans l'ensemble ont bien su circonscrire les aspects importants du sujet, mais aurait souhaité plus de précision et de rigueur notamment sur les parties ayant trait à la biologie cellulaire.

2. Forme et méthode

Rigueur, vocabulaire et rédaction

Le plan du développement doit être apparent et les titres des paragraphes doivent refléter la progression du plan, être précis et adaptés à leur contenu. Par ailleurs, la biologie possède un **vocabulaire précis** pour désigner chaque élément, structure, phénomène ou concept. Il est donc primordial de veiller à la **justesse du vocabulaire** utilisé (confusions entre brassages intra- et interchromosomiques, terme « non-codant » utilisé pour désigner des variations codantes synonymes, manque de distinction entre notions proches telles que mutation, allèle et polymorphisme, ou encore entre gène et allèle, etc.). Il est à noter que même si on ne peut que conseiller aux candidats de présenter des copies mieux rédigées, la syntaxe et l'orthographe ont été plutôt satisfaisantes.

Introduction et plan

L'introduction doit replacer le sujet dans un contexte plus large et l'explicitier pour arriver à dégager une problématique qui donne le fil directeur du développement et justifie le plan annoncé. Elle doit être composée : (i) d'une remise en contexte, (ii) d'une explicitation du sujet avec définition des termes importants, (iii) d'une délimitation précise et justifiée du sujet, (iv) d'une problématique et (v) de l'annonce du plan.

Le **cadre général** posé au début de l'introduction doit être plus large que celui du sujet. Il peut s'appuyer sur des exemples, des faits concrets ou théoriques (par exemple une pathologie liée à un variant génétique) et éveille la curiosité du lecteur. Il amène progressivement à définir les termes du sujet.

La **définition des termes** (ici Variabilité génétique et Population humaine) doit être précise et approfondie car elle conditionne la délimitation du sujet, c'est-à-dire ce qui sera traité ou non dans le développement. Ici, les notions à définir étaient celles de génome comme support de l'information génétique, d'individu porteur d'un génome qui lui est propre, et de population comme ensemble d'individus, permettant d'explicitier la notion de variabilité génétique comme l'ensemble des différences

généétiques entre individus au sein de la population. Par ailleurs, une définition générale de la population humaine était attendue : espèce eucaryote, appartenant aux vertébrés, mammifères et primates (le jury a valorisé toute information phylogénétique correcte), comprenant environ 7 milliards d'individus aujourd'hui. Malgré cela, trop peu de candidats ont fait l'effort de définir précisément ces termes et se sont contentés d'annoncer les termes de génome et de variation génétique sans les définir.

La **délimitation du sujet** doit expliciter les notions à inclure et exclure du sujet. Ici, l'apparition, la transmission et les conséquences de la variation génétique chez l'être humain étaient toutes à inclure, mais les mécanismes moléculaires généraux du fonctionnement du génome (réplication, transcription, traduction) n'étaient pas à expliciter en détail. Par ailleurs, le sujet ne portant que sur l'espèce humaine, les mécanismes et exemples spécifiques aux génomes procaryotes étaient hors sujet.

La **problématique** doit être exposée sous forme de question (une ou plusieurs) explicite à laquelle le développement s'efforcera de répondre. Elle doit donc envisager l'ensemble des aspects du sujet et soulever un problème qui sera le fil directeur du développement pour préparer l'annonce du plan qui suit. Une problématique du type « Pourquoi les êtres humains sont-ils différents ? » n'a que peu de sens et ne recouvre pas l'entièreté du sujet. De même une problématique du type « nous discuterons les différents aspects de la variabilité génétique chez l'être humain » ne permet pas de construire une progression logique. Un exemple approprié pourrait consister à se demander comment de petites différences au niveau du génome se traduisent par d'importantes différences de phénotypes entre les individus.

Enfin, l'**annonce du plan** permet de dresser les grandes lignes de la progression qui sera suivie dans le développement. Des plans basés sur les différents types de polymorphismes présents dans la population humaine, ou sur les différentes forces évolutives qui les ségrégent, manquent de cohérence et de logique et pénalisent les candidats qui se répètent tout au long du développement ou passent sous silence des liens importants.

Le jury rappelle donc que la réflexion préalable donnant lieu à la rédaction de l'introduction doit être particulièrement travaillée, de manière à ce que **la problématique comme le plan soient réfléchis** pour pouvoir proposer un discours cohérent et construit.

Fil conducteur et progression logique du raisonnement

L'objectif d'un devoir de synthèse est de produire un raisonnement qui par une progression logique et argumentée réponde au sujet proposé. Ainsi, tout élément exposé doit être très clairement relié au sujet, ainsi qu'aux éléments le précédant et le suivant. Il est donc primordial de soigner l'ordre dans lequel les différents concepts sont abordés et les transitions entre chacun d'eux. Le sujet faisant appel à des notions d'échelle très variées allant du mécanisme moléculaire aux organismes et jusqu'aux dynamiques populationnelles et évolutives, l'articulation logique y était particulièrement importante afin de motiver le raisonnement. Le jury invite également les candidats à ne pas dévier de leur ligne directrice, et à ne pas s'égarer dans des digressions sans lien avec le sujet ou anticipant une étape ultérieure du raisonnement. Pour gagner en cohérence dans le développement, les transitions gagneraient à être plus travaillées pour ne pas se résumer à « on a vu ceci et on va maintenant voir cela. »

Démarche scientifique et expérimentale

La démarche scientifique, hypothético-déductive, n'est que trop rarement utilisée. La mention explicite de l'être humain dans le sujet a mis en difficulté certains candidats, la majorité des mises en évidence au programme concernant d'autres espèces. Pourtant, de nombreuses mises en évidence expérimentales simples, basées sur l'observation, étaient possibles, notamment en utilisant des exemples tirés de la génétique classique chez l'homme. Par ailleurs, le jury a jugé appropriées toutes les mises en évidence se basant sur des espèces non humaines, quand leur pertinence en tant que modèle pour l'être humain a été explicitée (par exemple, en indiquant qu'il s'agit également d'une espèce eucaryote, ou d'un mammifère au génome proche du nôtre). La biologie est une science expérimentale qui s'appuie sur l'observation. Se limiter à un exposé théorique et une accumulation de notions et de modèles ne peut que diminuer l'impact des démonstrations. Il est donc important que la démarche expérimentale apparaisse dans cette épreuve ; elle est fortement valorisée par le jury.

Exemples et illustrations

Le jury félicite les candidats qui utilisent des exemples et illustrations pour étayer leur discours et leurs idées. Attention toutefois car un exemple peu précis ou juste évoqué ne permet souvent pas de soutenir un raisonnement.

Par ailleurs, des exemples chiffrés (ordre de grandeur) tels que la taille du génome humain, le nombre de substitutions de base par cycle de réplication, la fréquence des enjambements lors de la méiose, etc sont rarement énoncés alors qu'ils sont nécessaires pour rendre la démonstration précise et rigoureuse. Les figures sont globalement de bonne qualité. Nous rappelons qu'elles doivent comporter un titre et une légende explicite. Elles doivent être soignées et illustrer précisément la notion présentée. Il est souvent peu opportun de réaliser un schéma de cours sans l'adapter au sujet. Par exemple, un schéma complet et détaillé de la structure moléculaire de l'ADN n'était pas particulièrement pertinent ici : seules les bases et leur complémentarité sont nécessaires pour traiter le sujet. De même, pour matérialiser les différentes étapes d'un processus, l'utilisation de flèches ou de numéros est une solution souvent pertinente.

Conclusion

Toute conclusion doit comporter un bilan des idées essentielles développées pendant l'exposé et répondre à la problématique énoncée en introduction. Dans l'exemple énoncé plus haut, on pouvait ici décrire les mécanismes qui mènent des différences minimales de patrimoine génétique aux conséquences importantes aux échelles de l'organisme et de la population. Trop peu de conclusions y répondent et se contentent bien souvent de paraphraser les grandes idées énoncées en introduction sans mettre en exergue les avancées conceptuelles réalisées durant le développement. Enfin, la conclusion propose un élargissement du sujet pour dépasser les limites fixées en introduction ou proposer des applications. De nombreux candidats ont proposé des pistes d'ouverture avec des inspirations diverses, en particulier sur la médecine génomique et les modifications du génome par CRISPR, ce qui a été généralement apprécié par le jury.

ANALYSE DE DOCUMENTS

1. Remarques et conseils méthodologiques

L'analyse de document permet d'évaluer diverses compétences scientifiques :

- (1) la capacité à **analyser des résultats expérimentaux**, y compris **de manière critique**,
- (2) la capacité à **comprendre des protocoles** et à **proposer soi-même des expériences**,
- (3) la capacité à **réaliser le bilan d'une série d'expériences** se rapportant à un même sujet, à prendre du recul vis-à-vis d'elle et à en tirer des conclusions cohérentes,
- (4) les capacités à mener une **réflexion argumentée sur les phénomènes biologiques** mis en évidence et à **proposer des hypothèses pertinentes** permettant de les expliquer

Il s'ensuit que **la simple description des figures ne suffit pas** pour réussir ces exercices. On ne peut espérer comprendre les phénomènes biologiques étudiés dans ces documents en *n'analysant les figures que partiellement*, et en concluant sans prendre en compte l'ensemble des informations présentées. Passer d'une figure à l'autre sans jamais dépasser le stade de la description, comme c'est le cas dans beaucoup de copies, ne permet pas au jury d'évaluer les compétences mentionnées ci-dessus.

Par ailleurs, le jury attire l'attention sur la nécessité de **comprendre la question biologique et les expériences** décrites pour pouvoir exploiter les résultats obtenus, sans quoi l'analyse et l'interprétation des figures sont souvent partielles et erronées, donc inutiles. Il est donc indispensable pour les étudiants préparant le concours, et se destinant à une carrière scientifique, de **maîtriser la méthode d'étude de documents**, et de l'appliquer avec rigueur.

Le jury félicite les candidats qui ont fait preuve de persévérance dans la compréhension et l'élucidation des analyses de document proposées. Il déplore en revanche qu'un certain nombre de candidats aient choisi de parcourir l'ensemble des questions à la recherche de points « faciles » et aient abandonné les questions demandant de la réflexion, questions bien souvent clefs pour la compréhension du phénomène étudié. Cette stratégie ne permet pas de réussir une telle épreuve car le barème valorise fortement les compétences (2), (3) et (4) susmentionnées.

Enfin, nous rappelons ici que, comme dans toutes les épreuves du concours, les compétences de **communication scientifique** sont évaluées : il est attendu la construction d'une argumentation rigoureuse, suivant une progression logique et en distinguant précisément les liens de causalité, les simples corrélations et les hypothèses.

Dans le but d'aider les candidats se préparant à cette épreuve, le jury souhaite donc rappeler ici quelques conseils méthodologiques :

- **Lire attentivement l'énoncé.** L'énoncé comprend le texte, mais aussi les légendes des figures. Il contient les informations essentielles à la compréhension des documents et pose l'objectif de chaque expérience. Par exemple l'énoncé de la seconde partie indiquait que les humains modernes sont apparus en Afrique et se sont étendus en Europe et Asie où ils ont pu rencontrer les Néandertaliens.
- **Comprendre les expériences réalisées.** Lorsque le principe des expériences n'est pas supposé connu par les candidats, il est rappelé dans l'énoncé. Le candidat doit alors comprendre la question biologique posée et le principe de l'expérience réalisée avant de se lancer dans l'analyse des résultats.

- **Analyser** les documents **de manière détaillée** en s'appuyant sur la légende : il convient de comparer les conditions deux à deux, de décrire et quand c'est possible de quantifier les variations temporelles, de décrire précisément les images, d'analyser et de comparer les grandeurs des paramètres mesurés, de s'interroger sur la significativité des différences observées, enfin, de s'intéresser à l'ensemble des données présentées dans la figure.

- **Utiliser un vocabulaire précis et adapté.** Dire qu'une valeur est « normale » n'a pas la même signification que « la valeur est identique à celle trouvée dans le contrôle ». De même, dire que « UBR4 influe sur la mitose » n'est pas informatif. Le jury accorde une grande importance à la rigueur scientifique et au choix des termes utilisés. Il est difficilement compréhensible (et un peu inquiétant) qu'après deux ans de classe préparatoire, on puisse encore écrire que « l'homme moderne descend de Néandertal » (voire du chimpanzé) ou que « les néandertaliens n'étaient qu'au début de leur évolution et ils ne possédaient pas beaucoup d'attributs humains ».

- **Réaliser des schémas complets.** Tout schéma doit être légendé et explicité : un schéma non compréhensible par le jury ne pourra pas être valorisé.

- **Interpréter et conclure à l'issue de chaque figure.** En restant au niveau de la simple description ou de l'analyse superficielle, il est difficile de progresser dans la compréhension du phénomène biologique étudié. Il est nécessaire d'exploiter l'ensemble des données contenues dans le document, puisque toutes sont utiles pour appréhender le phénomène proposé. Le jury recommande de suivre toutes les étapes du schéma classique suivant : (1) analyse des résultats, (2) interprétation des résultats, (3) conclusions (notamment s'il y a plusieurs sous-figures) et éventuellement (4) énoncé d'hypothèses explicatives. Il est essentiel de commencer par l'analyse des témoins pour pouvoir conclure rigoureusement. Enfin, il ne faut pas sur-interpréter les résultats expérimentaux, et déduire des relations de cause à effet à partir de simples corrélations.

- **Proposer des hypothèses explicatives mais veiller à les différencier des conclusions (en les mentionnant au conditionnel notamment) et à les justifier.** Si le jury encourage les candidats à proposer des hypothèses après l'interprétation des résultats, celles-ci ne doivent pas être reprises dans les questions suivantes comme des éléments démontrés, au risque d'aiguiller l'analyse dans une mauvaise direction. Elles doivent au contraire être testées à nouveau et remises en question par les nouvelles données. De plus, lorsqu'il est demandé de proposer des hypothèses celle-ci doivent être justifiées sur la base d'arguments et de données issus des connaissances, des données de l'énoncé ou des conclusions tirées des résultats et non reposer sur une intuition.

Il est par ailleurs essentiel pour des étudiants engagés dans des études scientifiques de pouvoir réfléchir et envisager, pour chaque résultat, les différentes interprétations/hypothèses explicatives possibles. Il est parfois explicitement demandé de proposer plusieurs hypothèses (question 3a), mais parfois, ce sont aux candidats de proposer plusieurs hypothèses (question 5b, 9c).

- **Intégrer l'ensemble des données apportées par les différentes figures dans un modèle général qui doit être construit tout au long de l'exercice.** Les sujets sont conçus pour permettre aux candidats de proposer une vision d'ensemble du mécanisme étudié en faisant la synthèse de leurs interprétations successives. Sauf cas particulier (ou mentionné), les figures ne sont pas indépendantes

les unes des autres. Certaines questions, en général à la fin de chaque partie, incitent un gros effort de synthèse (questions 11 à 13) qui ne doit pas se limiter à résumer la question précédente. Le jury attend que les réponses aux différentes questions soient cohérentes entre elles et s'appuient sur (ou soient discutées en fonction) des réponses précédentes. Si des incohérences apparaissent, ce qui est parfaitement possible lors de raisonnements scientifiques, le jury valorise le fait qu'elles soient commentées de manière pertinente.

- **Maîtriser les techniques de base de biologie**, tant dans leur protocole que dans leur principe. Le principe du western blot et sa finalité sont mal maîtrisés tout comme le rôle du témoin de charge qui doit par ailleurs être analysé, et non pas simplement évoqué. Le vocabulaire spécifique des techniques est souvent aussi approximatif : par exemple, dans les western blot, on parle de bandes (et non pas de taches, traits, ou autre).

- **Concision, soin, clarté et rédaction.** Il est inutile d'introduire chaque expérience par une description du protocole qui paraphrase le texte. Il est préférable d'approfondir l'analyse. De plus, l'écriture et la syntaxe doivent être soignées de manière à être déchiffrable et compréhensible. Il est fortement recommandé de préférer des phrases courtes mais claires, à des phrases longues comportant de si nombreuses propositions subordonnées qu'elles en deviennent incompréhensibles.

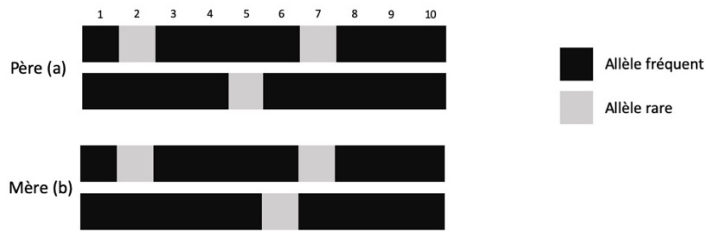
2. Eléments de correction de la partie B

Question 1 : Les SNP sont des substitutions dont les causes peuvent être exogènes (UV, produits chimiques etc) ou endogènes (erreurs de réplication ou de réparation par exemple). Sur la figure 1, l'oligonucléotide sert d'amorce pour l'incorporation d'un nucléotide fluorescent. Ces nucléotides étant di-déoxy, ils ne permettent pas l'incorporation d'autres nucléotides, et la mesure de la fluorescence permet donc d'identifier le nucléotide incorporé. Chaque individu étant diploïde, on détecte un signal de fluorescence de 1 pour les homozygotes ou de 0,5 pour les hétérozygotes.

Dans la population africaine la fluorescence détectée pour 19 individus est à 598 indiquant qu'ils sont homozygotes pour T. Un individu a incorporé autant de T que de C et il est hétérozygote T/C. Dans la population européenne, environ 1/3 des individus sont homozygotes C/C ; 1/3 sont homozygotes T/T et 1/3 sont hétérozygotes T/C.

Les fréquences alléliques devaient être calculées sur les 40 allèles de chaque population. Ainsi, dans la population africaine, on comptait 39 allèles T et 1 allèles C. Donc $f(T) = 39/40$ (0,975) et $f(C) = 1/40 = 0,025$. Dans la population européenne, les fréquences approximatives (on ne connaît pas exactement le nombre d'individus de chaque génotype) étaient de 0,5 pour les allèles C et T.

Question 2 : La question 2 permettait d'aborder la notion d'haplotype. Les allèles 2 et 7 se comportent de la même manière car les allèles rares sont portés par le même chromosome, comme indiqué sur le schéma suivant :



Dans cette région du génome, trois haplotypes sont donc présents dans cette famille (un haplotype est commun au père et à la mère).

La principale difficulté de cette approche est que les individus étant diploïdes, il est difficile de savoir si les allèles identifiés viennent du chromosome maternel ou paternel. De plus, au cours du génotypage, on perd l'information de liaison des locus.

Question 3 :

Trois hypothèses non exclusives permettent d'expliquer les différences de fréquences des trois haplotypes dans la population. On peut imaginer que ces haplotypes ont été sélectionnés sur la base d'avantages sélectifs qu'ils confèrent selon les environnements. Alternativement, on peut envisager que les haplotypes Ha, Hb et Hc sont apparus en Afrique, en Europe et en Asie respectivement et qu'ils sont restés fréquents dans leur région d'origine et plus rares ou absents ailleurs. Enfin, on peut proposer que pendant les migrations depuis l'Afrique vers l'Europe et l'Asie, certains haplotypes ont augmenté en fréquence par dérive génétique au sein de groupes de petite taille et d'autres se sont perdus. Si la première hypothèse a quelquefois été proposée, les autres ne l'ont été que de manière tout à fait exceptionnelle.

L'haplotype B correspond à un chromosome recombinant entre Hc et Ha, chez un individu sans doute d'Asie du Sud-Est ou éventuellement d'Afrique.

L'analyse d'une partie seulement des SNP est permise par le fait qu'ils sont souvent liés et on peut donc déduire les SNP les plus probables autour des SNP génotypés. Cette méthode est plus rapide et moins onéreuse mais est source d'erreurs en cas de recombinaison. On peut également noter, mais il n'était pas attendu des candidats qu'ils l'indiquent, qu'elle s'appuie sur les SNP les plus fréquents et fait donc perdre de l'information sur les SNP les plus rares.

Question 4 :

Dix SNP sont représentés chez les individus i, j et k et il y avait donc 20 SNP analysés (les individus sont diploïdes). On comptait 3, 4 et 5 différences entre j-k, j-i et i-k respectivement, ce qui équivaut à des distances génétiques de 0,15 (3/20) ; 0,2 (4/20) et 0,25 (5/20). Ces distances pouvaient être représentées par un triangle rectangle. La figure 4 montrait que la proximité génétique des individus reflète leur proximité géographique et les axes de variation principaux correspondent aux axes géographiques Nord-Sud et Est-Ouest.

Question 5 :

Les questions précédentes montraient que certains SNP ou haplotypes sont plus fréquents dans certaines régions du monde. En séquençant des génomes d'individus d'origines connues, on peut

associer des haplotypes à des régions géographiques probables. La précision du placement géographique dépend donc du nombre de génomes analysés et il est possible qu'elle soit plus grande en Europe ou aux États-Unis et plus petite dans d'autres régions du monde. D'autre part, l'analyse ne décrit pas une réalité absolue mais une simple probabilité d'origine.

Certaines régions ne sont pas assignées si elles contiennent des haplotypes/SNP dont la fréquence est la même dans toutes les régions du monde ou au contraire, si elles proviennent d'une région où trop peu d'individus ont été génotypés. On ne peut pas exclure non plus que pour des raisons techniques, certaines régions du génome ne soient pas analysées.

L'individu a un parent du Liban, et un autre dont l'origine est à environ 50% européenne et 25% asiatique. Ce parent a donc lui-même deux grands-parents d'origine européenne, un grand-parent d'origine asiatique et un quatrième grand-parent ni Asiatique ni Européen sans quoi il aurait été identifié. Certains candidats ont dessiné un arbre généalogique annoté pour répondre à cette question, une solution élégante qui a été très appréciée par le jury.

Les questions 6 à 9, portant sur l'identification des ADN néandertaliens ont été particulièrement peu réussies.

Question 6 :

Le faible pourcentage d'ADN d'origine humaine détecté sur la figure 6A peut s'expliquer par le fait que les os ont vraisemblablement été colonisés par des bactéries et des micro-organismes. Pour le savoir, on peut comparer les séquences non-humaines aux génomes des bactéries du sol. L'ADN humain, plus ancien a subi une oxydation des groupements amines des cytosines qui sont transformées en Uracile. Ces modifications ne sont détectées qu'aux extrémités ce qui suggère que les extrémités sont plus fragiles, ou plus exposées que le reste des fragments.

Question 7 :

Les expériences décrites dans la figure 7 montrent que l'ADN peut se complexer aux cristaux d'hydroxyapatite, donc aux os (et il se retrouve alors dans le culot) grâce à ses groupements phosphates, qui entrent en compétition avec les ions phosphates de l'hydroxyapatite des cristaux qu'on retrouve pour leur part dans le surnageant. La figure D montre que l'hypochlorite (l'eau de javel) dégrade l'ADN en solution, mais pas l'ADN complexé à l'os.

Question 8 :

En utilisant cette stratégie, on peut débarrasser les os fossiles des ADN contaminants, et ainsi augmenter fortement la proportion d'ADN humain (cf figure 6A). Dans les os récents, l'ADN humain est sans substitution puisqu'il n'a pas eu le temps d'être oxydé. La proportion d'ADN sans substitution, donc récent, diminue après traitement à l'hypochlorite pour les échantillons des grottes françaises et belges. Les échantillons de la grotte croate avaient peu ou pas de contamination par de l'ADN humain moderne. Les contaminations des os fossiles par de l'ADN humain moderne sont sans doute dues aux manipulations des scientifiques qui ont découvert les grottes. C'est pour cette raison que les extraits découverts plus tardivement (en 2017 en Croatie) n'ont plus ce type de contamination car des précautions ont été prises lors de l'excavation. Il n'était pas attendu des candidats qu'ils tiennent ce raisonnement.

Question 9 :

Les séquences néandertaliennes ont été intégrées dans le génome humain moderne du fait des croisements entre l'homme moderne et les néandertaliens, qui sont décrites dans l'énoncé et ont sans doute été sélectionnées (positivement ou négativement). Les critères utilisés pour retrouver les SNP néandertaliens permettent d'identifier 1-l'ancêtre commun et 2-l'allèle néandertalien. Le troisième critère s'appuie sur le fait que les Yorubas vivent en Afrique d'où est sorti l'Homme moderne, avant les croisements avec les néandertaliens (ces informations étaient données dans le paragraphe introductif). Des allèles présents chez les Yorubas et chez les néandertaliens sont donc probablement ancestraux. La figure 9A et B montrent que les pics et les zones sans allèles commun avec les néandertaliens sont aux mêmes locus dans les génomes européens et asiatiques. Cela suggère qu'il y a eu des pressions de sélection négatives et positives. Les allèles néandertaliens sont détectés à très faible fréquence dans la population africaine, ce qui est cohérent avec l'absence de croisement entre les ancêtres africains et les néandertaliens.

Le faible taux de séquences néandertaliennes sur le chromosome X peut s'expliquer par le fait que peu de chromosomes X néandertaliens ont été introduits dans la population humaine moderne, ce qui est le cas si les femmes issues de croisement mixtes sont stériles ou pas viables. Alternativement, il est envisageable que les séquences néandertaliennes sur le chromosome X ont été contre-sélectionnées. Dans le tableau 9B, il était attendu que les candidats commentent le fait que les populations modernes non africaines portent plus de 1% de séquences néandertaliennes. Il n'était pas attendu que les candidats le notent (et aucun candidat ne l'a noté), mais ce faible pourcentage représente quand même 30 millions de paires de base, ce qui est loin d'être négligeable. Les candidats pouvaient ensuite remarquer que comme attendu, ce taux de séquence néandertaliennes était quasiment nul dans les populations africaines, alors que le taux intermédiaire dans les populations américaines d'origine africaine reflète son histoire de métissages.

La longueur totale d'ADN néandertalien dans le génome actuel pouvait être calculée en additionnant les longueurs des fragments présents en fonction de leur distribution ($\sum_{i=1}^n F_i \times L_i$ où F_i est le nombre de fragments de taille L_i). Cette longueur représente environ $0,8 \times 10^9$ pb d'ADN. Environ 25% du génome néandertalien est donc conservé dans la population humaine actuelle, mais pour chaque génome, seulement 1% est d'origine néandertalienne. Ces chiffres ont été rarement trouvés et encore plus rarement commentés.

Question 10 :

Puisqu'on a séquencé le génome de chaque individu moderne dont on a mesuré la globularité crânienne, on peut comparer pour chaque position polymorphique le score moyen de globularité pour tous les individus qui ont l'allèle moderne et le score moyen pour tous les individus qui ont l'allèle néandertalien. Les positions pour lesquelles le score de globularité est plus petit pour les individus qui ont l'allèle néandertalien sont des gènes candidats.

Un nombre relativement important de copies ont proposé des stratégies d'inactivation géniques ou de transgénése, oubliant probablement qu'il s'agissait ici d'études sur l'humain. Quelques candidats ont cependant proposé ces solutions tout en mentionnant explicitement qu'elles n'étaient pas éthiquement envisageables.

Les **questions 11 à 13** demandaient une grosse capacité de d'analyse et de synthèse. De plus, beaucoup des données présentées dans les figures 12 et 13 permettaient d'émettre des hypothèses mais ne permettaient pas de formuler des conclusions fortes. Elles ont été discriminantes.

Des figures 11 et 12, il pouvait être proposé que chez les néandertaliens, PHL1 est plus fortement exprimé dans le cervelet qui est une des régions du cerveau les plus différentes entre les Hommes modernes et les néandertaliens. Dans cette région, la forte expression de PHL1 aurait pu induire de l'apoptose et donc diminuer la globularité du cerveau néandertalien, mais aussi diminuer la maturation des synapses, ou tuer les neurones dont les synapses sont matures.

Des figures 11 et 13, on constatait que UBR4 est moins exprimé dans le putamen des individus porteurs de l'allèle néandertalien. Chez la souris, la perturbation de l'interaction d'UBR4 avec Nde1, une protéine associée aux microtubules et impliquée dans le cycle cellulaire, modifie l'angle de division des précurseurs neuronaux et donc diminue la neurogenèse. Ainsi ce phénomène pourrait se retrouver chez les humains porteurs de l'allèle néandertalien, ce qui pourrait modifier la forme du cerveau et donc du crâne.